"13 Essays in Data Science"

On the intersection of Statistics, Software, and Atmospheric Sciences.

Habilitation Thesis

for obtaining the *venia docendi* in Data Science

presented to the
Faculty of Economics and Statistics
at the University of Innsbruck

by Mag. Reto Stauffer, PhD

Innsbruck, November 2024

Contents

Structure of the Habilitation

This cumulative habilitation thesis consists of thirteen scientific publications in accordance with the criteria for habilitations at the Faculty of Economics and Statistics of the University of Innsbruck as well as § 103 UG 2002.

The structure of the thesis is as follows: First, I will give a brief description of my academic career and how my fascination for programming and my curiosity of statistics shaped my path into data science. Second, an overview of the scientific contributions which make up this thesis is provided, followed by the individual publications.

According to the faculty criteria, each article has a rank assigned based on the Journal Citation Report (JCR, Clarivate Web of Science) using the five-year average impact factor. Articles published in journals with an average score below the median in their field are labeled as category 2, those above the median as category 1. Additionally, articles published in a selection of the five highest-ranked statistics journals are labeled as 'Top 5'. Overall, this thesis consists of one 'Top 5' article (Article V; published in the *Annals of Applied Statistics*), seven category 1 articles (Articles I, II, III, VIII, X, XII, and XIII; 2× *Statistics and Probability*, 5× *Meteorology & Atmospheric Sciences*), one category 2 article (Article IX; *Meteorology & Atmospheric Sciences*), as well as four articles published in recent peer-reviewed journals that are not yet ranked in JCR (Articles IV, VI, VII, and XI; 3× *Meteorology & Atmospheric Sciences*, 1× *Software*).

The contributions listed for publications where I am not the lead author are based on the contributor role taxonomy (CRT, National Information Standards Organization CRediT).

My Journey in Data Science

With a major and a PhD in atmospheric sciences, my original training focused on understanding the physics of the atmosphere (i.e., weather and climate). The expertise I gained during this period, combined with my passion for programming and statistics, has shaped my career path and defined the focus of this thesis. This interdisciplinary and transdisciplinary field of expertise is often referred to as data science. Although there is no universal definition of "data science", it is usually described as a combination of expertise in statistics, programming, and domain-specific knowledge.

I bring a strong educational background in atmospheric sciences, extensive expertise in applied statistics and probability, and a diverse skill set in software development, spanning various programming languages and environments. My experience includes utilizing high-performance computing, managing custom UNIX-based servers for a variety of services, and developing and maintaining a range of software packages and applications. This includes the migration, customization, development, and maintenance of the Journal of Statistical Software in my role as Technical Editor for the journal since 2015.

Since 2019, I have held a tenure-track position at the newly founded Digital Science Center (DiSC) at the University of Innsbruck, where I have made key contributions to the "Digital Science" minor program. This includes co-creating the foundational courses "Introduction to Programming" (in *R*) and "Introduction to Data Management", helping to advance digital science education.

The expertise I've acquired goes beyond my formal education and reflects my passion and dedication to data science, allowing me to always act proactively when new challenges, ideas, or problems arise. This commitment extends beyond my employment, as I continue to support colleagues, partners, students, and friends with the knowledge I have gained over the years—and beyond.

Overview of Scientific Contributions

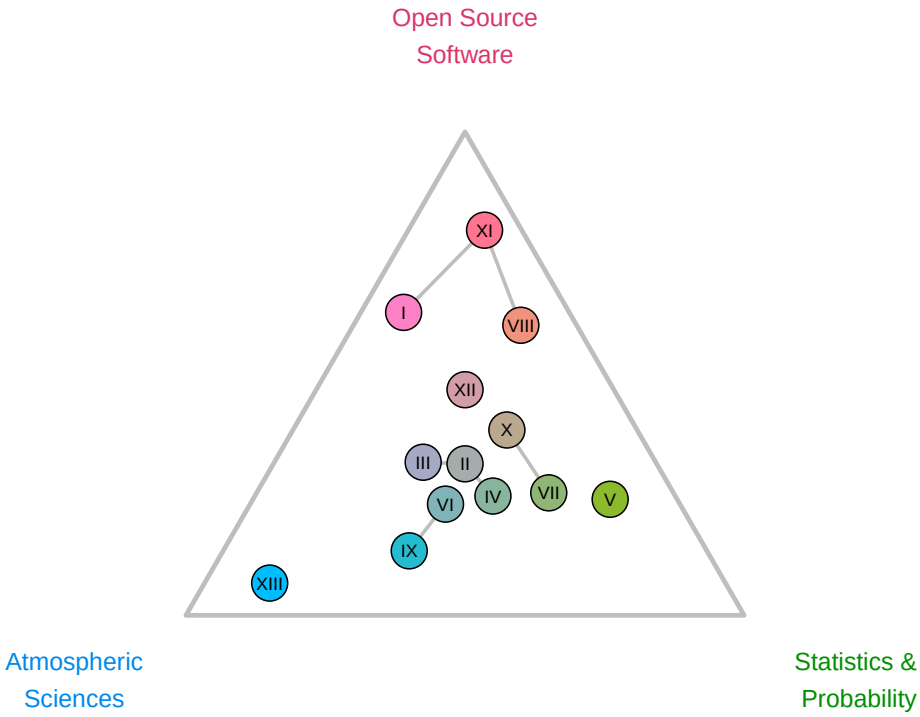

Figure 1: Graphical summary of the contributions in this habilitation thesis. Each article is represented by a marker, with its position indicating the focus across the three main areas of expertise. Articles near the center reflect a balanced focus across all three areas. Gray lines connecting articles denote groups of related articles, which are used as a basis for the outline on the following pages.
Article legend: Somewhere over the Rainbow (I), Spatio-Temporal Precipitation Climatology (II), Daily Precipitation Sums over Complex Terrain (III), Hourly Probabilistic Snow Forecasts (IV), Distributional Regression Forests (V), Skewed Distribution for Temperature Forecasts (VI), Bivariate Gaussian Wind Forecasts (VII), *R* colorspace (VIII), Time-Adaptive Training Schemes (IX), Circular Regression Trees and Forests (X), Python colorspace (XI), Long-Term Foehn Reconstruction (XII), Atmospheric Deserts (XIII).

Figure 1 provides a graphical summary of the thirteen peer-reviewed scientific articles "on the intersection of *Statistics*, *Software*, and *Atmospheric Sciences*" compiled in this thesis. Each article relates to some extent to all three topics, as represented by its proximity to the corners of the ternary diagram.

The following sections provide a summary of the publications that form the foundation of this thesis, with related contributions grouped together as indicated by the connecting gray lines in Figure 1.

The first section is dedicated to effective color palettes for (scientific) visualizations, followed by publications more closely related to atmospheric sciences and applied statistics. Since nine of the thirteen articles focus on statistical postprocessing in atmospheric sciences, a brief introduction to weather forecasting, statistical postprocessing, and distributional regression is included to define key terms for readers less familiar with this field.

Effective color maps for (scientific) visualizations (I, VIII, XI)

Color is an integral element of visualizations and graphics, often used for communicating (scientific) results, findings, or information in general. For a long time many software packages have used color palettes from simple red-green-blue (RGB) color combinations, the most notable example being the (in-)famous RGB "rainbow" or "jet" color palette. However, poor color palette choices can create various problems, ranging from simply being visually unappealing to obscuring key information or making graphics unreadable for audiences with color vision deficiencies.

Our contribution over the past years aim to raise awareness among different communities, and to provide guidance on how to prevent common problems. Moreover, our goal is, was, and will continue to be to equip everyone–from novices to developers to experts–with tools and software to facilitate choosing, customizing, assessing, and implementing effective color palettes in their workflows and projects in *R*, Python, and beyond.

Stauffer et al. (2015, I) specifically target a broad audience of meteorologists and atmospheric and climate scientists. Due to a combination of historical reasons and established habits, inefficient color palettes were used widely in this field. Our article, published in the Bulletin of the American Meteorological Society, reached a large readership and helped to raise awareness of this important topic, sharing the same goal and complemented other initiatives at that time, such

as "The end of the rainbow" by Hawkins et al. (2014) with an open letter to the climate science community. This article has been published before some of the widely known graphical libraries changed to better defaults, including Python's `matplotlib` (Hunter et al. 2017) and base *R*, which changed their defaults with the release of version 2.0 (Hunter et al. 2017) and version 4.0.0 (Zeileis and Murrell 2023), respectively.

Zeileis et al. (2020, VIII) present all key features and capabilities of version 2.0.0 of the *R* package `colorspace`, which greatly benefited from community feedback and the efforts of its contributors over the years prior to publication. Since the package's early release, many features have been extended, refined, and added. Highlights include enhanced functions for accessibility checks, tools for color palette assessment, a new graphical user interface with multiple apps, full integration with the `ggplot2` library (Wickham 2016), and a range of advanced color palettes and presets. This work also contributed to the update of standard color palettes in base *R*.

To expand access to our suite of tools, we released the first major version of the Python `colorspace` package in 2024, providing access to all features to the Python community (native Python implementation). Stauffer and Zeileis (2024, XI) present an overview of the software, making `colorspace` accessible to a broader user base with the aim of sensitizing more users to this important aspect of visualization, as well as further improving and extending `colorspace`.

In addition to the peer-reviewed contributions summarized above, we have continuously aimed to reach as many users as possible, allowing everyone to make use of our work. Since 2015 (alongside Stauffer et al. 2015), we have made our tools available via our website, `https://hclwizard.org/`. The HCLwizard enables anyone with a web browser to use our apps and export colors and color palettes for use across various programming languages and use cases, not limited to *R* and Python.

Software contribution: Major updates, extensions, and refinements to the *R* package *colorspace*, as well as a fully native Python implementation (as lead developer) and the HCLwizard web platform, providing easy access to our suite of apps.

A brief introduction to statistical postprocessing

Weather forecasts are typically produced by physically-based numerical weather prediction (NWP) models. These models use vast amounts of observational data from different sources such as land stations, ships, buoys, and satellites to calculate the current state of the atmosphere (analysis). This serves as the initialization for the NWP model, which then solves a set of prognostic equations to simulate the future state of the atmosphere (forecast).

However, due to the availability and quality of observations, numerical approximations, necessary simplifications, and the atmosphere's inherent chaotic nature, these forecasts are never fully exact. To account for this uncertainty, multiple NWP forecasts are generated with slightly perturbed initial conditions and model formulations, creating an ensemble of forecasts. These systems known as ensemble prediction systems (EPSs) are used operationally since more than thirty years (e.g., Buizza and David 2017), and typically consist of 20 to 50 members (NWP runs) depending on the product. Although EPSs provide valuable uncertainty information, they often contain systematic errors and are known to be underdispersive, meaning that the uncertainty is often underestimated.

One approach to improving these forecasts is through statistical methods employing historical observations from weather stations along with past forecasts from an EPS to identify and correct systematic errors. In its simplest form this can be achieved using (multiple) linear regression:

$$y \sim \mathcal{N}(\mu, \sigma) \quad \text{with} \quad \mu = \beta \mathbf{X} \quad \text{and} \quad \sigma = \text{sd}(\epsilon), \tag{1}$$

where the response $\mathbf{y}$ is assumed to follow a Gaussian distribution $\mathcal{N}$ with location μ and standard deviation σ. In this most simple form, only μ is modeled by a linear additive term ($\beta \mathbf{X}$), where $\mathbf{X}$ includes a series of covariates or explanatory variables derived from the NWP output. Models of this type are often referred to as 'model output statistics' (Glahn and Lowry 1972) within meteorology, hydrology, and climate sciences.

To incorporate the uncertainty information provided by the EPS, these models can be further extended to:

$$y \sim \mathcal{N}(\mu, \log(\sigma)) \quad \text{with} \quad \mu = \beta \mathbf{X} \quad \text{and} \quad \log(\sigma) = \gamma \mathbf{Z}, \tag{2}$$

where the scale parameter σ is modeled by its own linear predictor $\gamma \mathbf{Z}$. Typically, the ensemble mean serves as covariate in $\mathbf{X}$, while the ensemble standard deviation or variance is used in $\mathbf{Z}$. This type of model, first proposed by Gneiting

et al. (2005), is often referred to as 'ensemble model output statistics' (EMOS) or 'non-homogeneous Gaussian regression' (NGR).

This framework can further be extended and generalized to what is known as 'generalized additive models for location, scale, and shape' (GAMLSS) or 'distributional regression' (see e.g., Rigby and Stasinopoulos 2005; Klein et al. 2015; Umlauf et al. 2018). In its general form, distributional regression is expressed as:

$$g_Y(\mathbf{Y}) \sim \mathcal{D}(h_1(\theta_1) = \eta_1, h_2(\theta_2) = \eta_2, \ldots, h_K(\theta_K) = \eta_K) \quad \text{with} \quad \mu_\bullet = \beta_\bullet \mathbf{X}_\bullet, \quad (3)$$

where the (potentially transformed) response $\mathbf{Y}$ is assumed to follow any distribution $\mathcal{D}$ from the exponential family, which can be either univariate or multivariate. Each parameter of the distribution $(\theta_1, \ldots, \theta_K)$ is modeled by its own linear predictor $(\mu_1, \ldots, \mu_K)$, allowing for linear, non-linear, and even multi-dimensional effects. Link functions $(h_1, \ldots, h_K)$ are applied to enforce certain constraints, such as ensuring positivity.

Over the past years, we have built upon this general approach by refining and applying existing methods, as well as developing new models to address specific research questions. This often involved the development of new software and/or modifications to existing software packages, ensuring that our research benefited from the open-source software community, and vice versa.

The following sections provide an outline of the scientific articles presented in this thesis (Fig. 1), focusing on key contributions and innovations. These advancements in statistical postprocessing and software development bridge the gap between theory and application in atmospheric sciences, advancing the field of data science.

Probabilistic snow forecasts over complex terrain (II, III, IV)

Living in the heart of the European Alps, where snow plays a key role in the local economy, strong snowfall events can significantly impact people's daily lives. The goal of the first group of contributions was to produce reliable probabilistic snow forecasts for Tyrol. This was achieved through a novel hybrid statistical postprocessing approach that integrates high-resolution standardized anomalies combining data with very different spatial and temporal extents.

Stauffer et al. (2017; II) introduce a new method to estimate a high-resolution spatio-temporal climatology of daily precipitation sums using distributional regression. To account for the nature of the data, a zero left-censored distribution is employed. The censored part accounts for the large fraction of zero observations (no precipitation), while an additional power transformation allows us to accurately model positive values. This novel approach enables the creation of high-resolution spatio-temporal estimates of the full climatological distribution.

This climatology serves as input for the spatial ensemble postprocessing of daily precipitation sums presented in Stauffer et al. (2017b, III), which for the first time use spatial standardized anomalies with additional censoring. Following the concept of Dabernig et al. (2017), climatologies were used to calculate standardized anomalies to remove all site-specific characteristics from the data, allowing to estimate a single regression model valid for the entire study area. This new approach enables accurate high-resolution probabilistic precipitation forecasts for any location within the area of interest.

Stauffer et al. (2017) and Stauffer et al. (2017b) culminated in Stauffer et al. (2018, IV), which introduces our novel hybrid approach for high-resolution probabilistic snow forecasts over complex terrain. The main challenges include combining data from sources with different spatial and temporal extents and accounting for conditions that lead to snowfall. Therefore, an additional model for probabilistic near-surface temperature forecasts is developed, serving as a proxy to distinguish between snow and rain events. To account for the correlation of these two quantities, ensemble copula coupling (ECC; Schefzik et al. 2013) is employed to restore the rank order structure of the EPS forecasts. The final result is reliable hourly high-resolution probabilistic snow forecasts that preserve the physical spatio-temporal coherence.

Software contribution: This research is mainly based on the *R* package `bamlss` (Umlauf et al. 2021), and necessary extensions and modifications required to conduct this research were incorporated directly into the package.

Distributional regression forests (V)

An alternative to regression-based supervised learners, as outlined in Equations 1–3, is the use of regression trees (Breiman et al. 1984). Unlike distributional regression models, where the relationships between covariates and the response are modeled as linear or smooth nonlinear effects, regression trees recursively split the data into more homogeneous subgroups. This approach allows capturing abrupt shifts in the data, approximating step functions to model highly nonlinear dependencies. An extension of regression trees is random forests (Breiman 2001), where an ensemble of trees is trained on resampled versions of the training data and then averaged. Random forests stabilize the partitions created by individual trees, offering better approximation of smooth functions and performing automatic variable selection by only splitting on informative covariates, a valuable aspect of this method.

However, classical regression trees and random forests model only the mean of the response, not the full distribution. A natural extension, therefore, is to adapt the framework to allow for splits that account for the full distribution of the data.

Schlosser et al. (2019, V) present a new generic framework for distributional regression trees and forests, where each split considers all parameters of the response distribution, dividing the data into subgroups that share similar distributional characteristics. To demonstrate the predictive performance of this new statistical method, probabilistic precipitation forecasts are produced for 95 stations across Tyrol (Western Austria) and its surrounding areas, using 80 covariates derived from an EPS. The results of this novel approach are compared to more traditional distributional regression models (EMOS, GAMLSS). The new method proves to perform on par with or even outperforms the traditional methods while only requiring minimal prior knowledge of the data, promoting itself as an attractive alternative to distributional regression.

Software: Implemented in the *R* package `disttree` (Schlosser et al. 2023) based on the `partykit` package (Hothorn and Zeileis 2015).

Seasonal varying coefficients and training schemes (VI, IX)

The two contributions summarized in this section address specific aspects with respect to probabilistic ensemble postprocessing for near-surface temperature. The near-surface temperature (air temperature 2 m above ground) is an often used quantity to develop, test, and validate new post-processing approaches, including the original work by Gneiting et al. (2005) introducing EMOS/NGR.

In this context, Gebetsberger et al. (2019, VI) present an extensive case study comparing various state-of-the-art EMOS models and distributional regression models. These model variations include multiple response distributions, introducing for the first time in this context the skewed logistic distribution (generalized logistic type I distribution) to address potential residual skewness. This study also explores different training schemes (seasonally varying regression coefficients vs. more classical training schemes) across different topographical environments, with a particular focus on forecast reliability and sharpness from multiple perspectives.

Lang et al. (2020, IX) build on one specific aspect of Gebetsberger et al. (2019), extending and deepening the analysis of the effects of different training schemes. Different researchers and teams working on the improvement of statistical postprocessing methods often use different training schemes. One common approach is the "sliding window approach", where models are trained solely on data from the most recent 30 to 60 days to allow the model to adjust for the current season. However, this approach limits the amount of training data. We proposed an alternative that leverages all available data for training while accounting for seasonally varying coefficients by incorporating cyclic smooth effects based on the day of year in the model specification. This article, for the first time, provides a detailed comparison of four different training strategies, examining their respective advantages and disadvantages in a case study on probabilistic temperature forecasting.

Software: This research is mainly based on the *R* packages `bamlss` (Umlauf et al. 2021) and `crch` (Messner et al. 2016). The additional family required to conduct the research presented in Article VI was incorporated in `bamlss`.

Probabilistic wind forecasting (VII, X)

This section provides an overview of two publications on advances in probabilistic wind forecasting. Wind is challenging due to its three-dimensional vector nature, though for near-surface applications, the vertical component can often be neglected as it is limited by the Earth's surface. Without this vertical aspect, wind can be described either as two vector components (u, v) (u: West-to-East; v: South-to-North), or in radial coordinates defined by direction and speed (magnitude).

Lang et al. (2019, VII) introduce a new probabilistic postprocessing method for wind forecasting using a distributional regression model. The two horizontal components (u, v) are assumed to follow a bivariate Gaussian distribution, where the location and scale parameters of both components are modeled by linear predictors. To account for dependence between u and v, an additional linear predictor is used to model their correlation, employing a rhogit link to ensure the correlation remains within $[-1, 1]$. This new approach not only corrects for biases or underdispersion in each component, but allows the model to make smooth rotation adjustments to correct wind direction misalignments in the ensemble forecasts. This is especially relevant for locations in complex terrain, where the EPS cannot fully resolve local topographic effects like terrain-induced channeling within narrow valleys.

For applications like airports, accurate short-term forecasts of wind direction changes can be of greater importance than the wind speed itself, as the wind direction defines the direction of airplane arrivals and departures. A change in wind direction may require to divert incoming and outgoing planes to a different runway. Thus, Lang et al. (2020, X) propose a new short-term wind direction forecasting method using a circular response distribution. The wind direction is modeled using the von Mises distribution, a two-parameter distribution that captures both location (main direction) and concentration (uncertainty). To allow capturing abrupt changes in wind direction, distributional forests are used (see Article V). The new method has proven to outperform other state-of-the-art circular GLM-type models for wind direction forecasting. Although the application in this article is related to a meteorological use case, the method presented expands the toolbox for modeling circular data beyond atmospheric sciences.

Software contributions: Distributional trees and forests are implemented in the *R* package `circtree` (Lang et al. 2024) based on the `partykit` package (Hothorn and Zeileis 2015).

Combining Supervised and Unsupervised Learning (XII)

The University of Innsbruck has a long history of foehn research due to its location in the heart of the European Alps where foehn and its effects can be experienced regularly. Foehn is a downslope wind on the leeward side of mountains, often characterized by a sharp increase in wind speed and changes in temperature and relative humidity. These characteristics allow for the detection of foehn using high-resolution measurements from meteorological stations. However, this limits detection to periods when such data is available, making long records of foehn very rare.

Stauffer et al. (2024, XII) present a novel approach that allows the reconstruction of long-term foehn time series spanning several decades. This is achieved by combining unsupervised and supervised learning, applied to the ERA5 reanalysis dataset, which provides atmospheric conditions at an hourly temporal resolution back to the year 1940. First, a two-component Gaussian mixture model with concomitants is used for foehn classification, utilizing 10-minute observations from meteorological stations. Second, this classification is used to train a series of binary response models (logistic regression, XGBoost) using information from ERA5. Once these models are estimated, they can be applied to the entire period from 1940 to 2022 to create the final reconstruction.

This novel dataset allows for an in-depth analysis of not only diurnal and seasonal patterns but also the detection of long-term trends with respect to the changing climate. For Altdorf in Switzerland, a unique and direct comparison between our reconstruction and an existing long-term foehn time series was possible. Although our approach has not seen most of the data from that period, the reconstruction shows a high degree of agreement with the existing dataset, demonstrating the accuracy and potential of this novel method.

Software contributions: The foehn classification was implemented in the *R* package `foehnix`, "A Toolbox for Automated Foehn Classification based on Mixture Models" (Stauffer 2023) available on GitHub. A Python implementation is available as well (see Dusch 2019).

Atmospheric Deserts (XIII)

Fix et al. (2024, XIII) introduce a new concept in atmospheric sciences, which we called "atmospheric deserts" (ADs), inspired by the concept of "atmospheric rivers" (ARs, Zhu and Newell 1998). In contrast to ARs, which are narrow atmospheric bands transporting large amounts of water vapor, ADs are air masses that are advected from hot and dry convective boundary layers in semi-arid or desert regions. ADs influence weather patterns by eliminating clouds, building up heat in the target area, and affecting the formation and suppression of thunderstorms. The article presents a new direct detection method for tracking ADs from their origin to destination using high-resolution Lagrangian trajectories.

We present a case study detecting and analyzing ADs originating from North Africa and traveling towards and across Europe. Using k-means clustering, four typical pathways of the calculated trajectories are identified, revealing varied thermodynamic evolutions influenced by factors such as condensation and radiative cooling. This first study helps to understand how ADs evolve along their path and how they influence local weather in the target region, paving the way for further research to enhance our understanding.

Article I

Stauffer R., Mayr G.J., Dabernig M., and Zeileis A. (2015). *Somewhere over the Rainbow: How to Make Effective Use of Colors in Meteorological Visualizations*. Bulletin of the American Meteorological Society, 96, 203–216, doi:10.1175/BAMS-D-13-00155.1.

JCR ranking: **Category 1** in *Meteorology & Atmospheric Sciences*.

SOMEWHERE OVER THE RAINBOW

How to Make Effective Use of Colors in Meteorological Visualizations

BY RETO STAUFFER, GEORG J. MAYR, MARKUS DABERNIG, AND ACHIM ZEILEIS

This paper offers a perception-based color space alternative to the well-known red–green–blue (RGB) color space and several tools to more effectively convey graphical information to viewers.

One of the many challenges associated with atmospheric sciences is the analysis and utilization of large, usually very complex datasets. One way to gather the information and to better understand it is to visualize it graphically. Visualizations may be as simple as one-dimensional plots (e.g., time series plots) or as complex as multidimensional charts (e.g., from numerical weather prediction model output). As a scientist, an important part of daily work is to create plots and graphs that visualize results and outcomes earned through weeks and possibly months of work. The key feature of visualization is helping the reader to capture the information as simply and quickly as possible. This reader can be a colleague, a customer, or even you.

AFFILIATIONS: STAUFFER, MAYR, AND DABERNIG—Institute of Meteorology and Geophysics, University of Innsbruck, Innsbruck, Austria; ZEILEIS—Department of Statistics, Faculty of Economics and Statistics, University of Innsbruck, Innsbruck, Austria
CORRESPONDING AUTHOR: Reto Stauffer, Institute of Meteorology and Geophysics, University of Innsbruck, Innrain 52, A–6020 Innsbruck, Austria
E-mail: reto.stauffer@uibk.ac.at

The abstract for this article can be found in this issue, following the table of contents.
DOI:10.1175/BAMS-D-13-00155.1

In final form 12 June 2014

The term "visualization" encompasses many aspects. Much work has been carried out during the last century to investigate the human perception and the influence of different aspects on how to best convey information. Initially, fundamental research was done in physics, biology, and medicine (see Miles 1943; Stevens 1966; Carswell and Wickens 1990), but with the advent of the computer industry this focus expanded into how to deal with the new technological achievements in different disciplines, including three-dimensional (3D) graphics, interactive visualization, and animation (Smith 1978; Ware 1988; BAMS 1993; Rogowitz and Treinish 1996; Light and Bartlein 2004; Hagh-Shenas et al. 2007). Today, the ubiquitous availability of computers and software enables everyone to create graphics for all different devices. We will focus on only one aspect: how to make effective use of color for visualization. Therefore, we are using relatively "simple" spatial plots to illustrate the guidelines and typical user tasks in atmospheric science.

Color is a good instrument to improve graphics, but carelessly applied color schemes can result in figures that are less effective than grayscale ones (Light and Bartlein 2004). For large parts of our vision, hue is irrelevant in comparison to shading. Color does not help us measure distances, discern shapes, detect motion, or identify small objects over long distances. Hue is useful for labeling and categorization but less effective for representing

(fine) spatial data or shape. However, if used effectively, colors are a powerful tool to improve (highly) complex visualizations. Therefore, it is important to know how color perception works and how we can make use of it to improve visualization (Ware 2004). Most common software packages supply methods to create different types of plots with different color maps (or color palettes). Nevertheless, because most plotting functions are rather generic it is impossible for the software developers to provide adequate color schemes for all applications.

The default color map is often a red–green–blue (RGB) rainbow palette. This is probably the most known color map and consequently many people use it uncritically as the default for their visualization, even though it has been shown to be difficult or even harmful (Brewer 1997; Borland and Taylor 2007). In addition to the rainbow scheme, other color maps defined in the RGB color space also exhibit similar problems and have to be handled with care, because the RGB color space has some critical disadvantages (Rogowitz and Treinish 1998; Light and Bartlein 2004). Vivid colors along the spectrum of the RGB color space strongly differ in their luminance, which can lead to artificial dark or bright bands that can obscure the information shown. Furthermore, the RGB color space is not a uniform color space, meaning that color pairs with the same distance within the color space do not show the same perceptual difference. Other prevalent color spaces, such as hue–saturation value (HSV) and hue–saturation–luminance (HSL), suffer from the same problems as the RGB color space (Smith 1978).

To avoid these disadvantages several transformations of the RGB color space have been developed. These transformations are approximations of the human color perception system, in that they, for example, allow for the fact that we have a logarithmic perception of luminance (Stevens 1966). The most profound work has been done by the Comission Internationale de l'Éclairage (CIE) in creating color spaces like the CIELUV or CIELAB color space, based on a standard observer (Ware 2004). Keep in mind that there is not "one omnipotent best" color model or color scheme. Depending on the user task or the medium, the most effective color models and palettes can differ. Even with extensive user testing there is always a number of different effective color maps for a given purpose. The work of Mahy et al. (1994) contains a good experiment-based comparison of many of the available color spaces. In this paper, we will focus on a perception-based color concept called hue–chroma–luminance (HCL), which is the CIELUV gamut in polar coordinates. Thus, the HCL color space is based on how humans perceive color, in contrast to the RGB color space, which is based on technical demands of TV and computer screens.

In this article, we will demonstrate the benefits of the HCL alternative, which is already becoming better known and more frequently used in other

HUMAN COLOR PERCEPTION

To choose the "best" (or at least a good) color scheme, one has to understand the characteristics of the receiver and processor. The human eye contains two classes of cells that are responsible for our visual perception: rod and cone cells. Rod cells serve vision at low luminance levels while cone cells are wavelength-sensitive. Three subclasses of cone cells are responsible for long, medium, and short wavelengths, respectively. Although most properly referred to as L, M, and S cones, respectively, the names R, G, and B cells are also frequently used, albeit somewhat misleadingly. The RGB annotation suggests that the cells refer to red, green, and blue, which is not the case. In fact, the LMS cones have broadly overlapping scopes. This design strongly differs from the "color separation" often built into physical imaging systems (Fairchild 2013b).

Under low luminance conditions, our visual system is limited to gather our surroundings by capturing differences in luminance using the rod cells only (monochromatic; scoptic view). Under moderate to high luminance conditions, the rod cells are fully saturated and our visual systems switch from a rod to a cone view (trichromat; photopic view). The spectral sensitivity also changes between these two views. Rod cells are more sensitive to shorter wavelengths. The appearance of two objects—say red and blue—changes under different luminance conditions. While they have the same lightness under high luminance conditions, the red one seems to look nearly black under low luminance conditions while the blue still looks quite light. This is caused by a lack of sensitivity to longer wavelengths (red) in the scoptic view. As one can see, the human perception is a complex system with different behaviors. All of these factors make it difficult, if not impossible, to represent colors that can be accurately perceived by humans in all settings/contexts.

Current theories describe that at least three dimensions are necessary to code a specific color. Typically, color models with three dimensions are employed (Knoblauch 2002), while one can argue that more dimensions would be required. For example, Fairchild (2013a) describes that five perceptual dimensions are necessary for a complete specification (brightness, lightness, colorfulness, saturation, and hue). However, typically only the relative appearance of the colors is of interest and not all five dimensions have to be known. Hence, the three perceptual dimensions—hue, chroma, and luminance—are typically sufficient for most purposes.

COLOR MAP DEFINITION

Most people are familiar with the coordinates of the RGB color space. Each of the three dimensions (red, green, and blue) can vary within the range from 0 up to 255. Table SB1 shows the coordinates for the HCL color maps used in this article. Choosing colors in the HCL space is similar to the RGB space, only the dimensions are different. The hue dimension (dominant wavelength) is circular starting with red (0), over green (120), to blue (240), and back to red (360 ≡ 0). The second dimension defines the chroma (colorfulness) and goes from 0 to 100. A chroma of 0 reduces the resulting color to a pure gray. The last dimension is luminance (brightness), which is also definable between 0 (black) and 100 (white). The transition between two values in each dimension should be monotonic but does not have to be linear (depends on the user task). For example, the luminance decrease in Figs. 4c,e follows a power function that allows us to highlight the areas with higher precipitation amounts (original value spacing in Fig. 4 is not linear either).

TABLE SB1. The coordinates for the HCL color maps used to create Figs. 1, 4, 5, and 6. The second column indicates the color followed by the coordinates for hue, chroma, and luminance. The first color indicates either the most left color (for horizontal color bars) or the bottom color (for vertical color bars). For the diverging scheme (Fig. 6b), the hue of the center value does not matter. It is exactly at the border for the two opposite hues, but its chroma is zero (gray).

Figure	Color	H	C	L
Fig. 1: Hue	1st	**0**	55	75
		*	*	*
	Last	**260**	55	75
Fig. 1: Chroma	1st	265	**0**	50
		*	*	*
	Last	265	**100**	50
Fig. 1: Luminance	1st	265	50	**15**
		*	*	*
	Last	265	50	**80**
Fig. 4c**	1st	80	20	92
		*	*	*
	Last	230	65	37
Fig. 4e**				92
	1st–4th	80	20	*
	5th–7th	130	35	*
	8th–10th	180	50	*
	11th–last	230	65	*
				37
Fig. 5d	1st (green)	150	76	98
	2nd (yellow)	92	82	81
	3rd (orange)	34	88	64
	4th (red)	–24	94	47
	5th (magenta)	–82	100	30
Fig. 6b**	1st (blue)	253	100	55
	Center (neutral)	—	0	95
	Last (red)	0	100	55

* Monotonically change along the dimension.

** Monotonic changes are nonlinear.

scientific fields (Zeileis et al. 2009; Silva et al. 2011). We argue that the use of misleading and distorting RGB color maps is not necessary, as alternative models are available and that changing to a perception-based color model can strongly improve the visual reception on graphical information with very little additional effort.

RGB VERSUS HCL COLOR SPACE. *HCL model.* The HCL color space is the polar transformation of the uniform CIELUV color space and forms a distorted double cone where each of the three dimensions directly controls one of the three major perceptual dimensions directly (additional information in the sidebars). The first one is hue, the dominant wavelength (defining the color); the second dimension is chroma, capturing colorfulness (color intensity compared to gray); and the third is luminance, pertaining to brightness ("amount" of gray). Figure 1 shows the three perceptual HCL dimensions. In each of the panels, one dimension changes linearly across the corresponding axis while the others are held constant. Hue changes the color while fixing the lightness and chroma level across colors. Increasing the chroma dimension increases the colorfulness compared to gray and the luminance dimension changes the colors from dark to light. Based on this concept, one can quickly define perception-based color maps for all kinds of tasks. Figure 2 shows a sample for HCL-based color maps. The qualitative schemes in Fig. 2a

are constant in chroma and luminance (passing from H_1CL to H_2CL), leading to isoluminant color schemes. Figure 2b shows sequential schemes with constant hue but increasing luminance and chroma (passing from HC_1L_1 to HC_2L_2), leading to a color map with a perceptual linear change. If hue is not constant along the color map, we will get multihue sequential schemes (Fig. 2b). The multihue color maps are similar to the single-hue sequences but here all three dimensions are changing from left to right (starting at $H_1C_1L_1$ and ending at $H_2C_2L_2$). The last examples (Fig. 1d) are of diverging schemes. Such color maps work well for data with two extremes around a neutral center value. Diverging color maps combine qualitative schemes (each side has a specific hue) with single-hue sequential schemes. The center value has high luminance and low chroma (leading to gray/white) followed by a symmetric and monotone increase in luminance and chroma. Because the different color spaces are connected by some coordinate transformation functions, each color can be expressed in coordinates of any other color space. This allows us to pick colors in the HCL space and convert them into the RGB color space, with which all software packages can deal.

RGB model. Historically, the RGB model is based on how screens work. Cathode ray tube (CRT), light-emitting diode (LED), and plasma screens attached to TVs, computer monitors, and projectors all use the same technique: Images are created by additive color mixing. Each image consists of hundreds to thousands of pixels where each pixel emits a mixture of red, green, and blue light. Each single RGB color is defined by a triplet of intensities for those three primary colors. Appropriate mixing produces a wide range of colors. Three zero intensities result in black, while maximum intensities for all three primary colors yield white and, in between, all other colors can be defined. Two widespread simple transformations of the RGB color space are the HSV (hexcone model) and the HSL (triangle model). Although they have a slightly better behavior, the basic problems of the RGB color space cannot be solved.

Desaturation. To focus on the luminance dimension of a color palette, it can be desaturated, for example, by transforming to HCL space, removing all chroma (so that hue does not matter), and transforming back to the original color space. This just removes hue/chroma information but keeps luminance fixed. In HCL dimensions, changes in hue or chroma do not influence the underlying luminance information.

Comparison of HCL and RGB. Figure 3 shows a juxtaposition of the (in)famous RGB rainbow color map and an alternative HCL rainbow. Both rainbows go from red over green and blue back to red. Below the color wheel, the same color maps are shown as colorized and desaturated color bars, respectively. As shown in the desaturated version of the RGB rainbow, even the three primary colors, red, green, and blue, vary enormously in luminance: red has a luminance value of about 50, green is about 86, and blue is about 30 (100 would be white). This creates unwanted gradients throughout the whole RGB rainbow map. Furthermore, the RGB color map shows several artificial narrow bands most easily seen around yellow/cyan/magenta. What makes the RGB color spaces even worse is that the different colors of the spectrum are not uniformly distributed (the green sector looks wider than the red one), which creates an additional distortion. In contrast, the HCL version shows an isoluminant gray in the desaturated version. This is no surprise, because one of the three dimensions of the HCL color space directly controls the luminance. Nevertheless, many common software packages provide the RGB rainbow scheme as default. As Borland and Taylor (2007, 14) wrote, "the rainbow color map is prevalent in the visualization community" even if "the rainbow color map [is often] a poor choice." Because of its

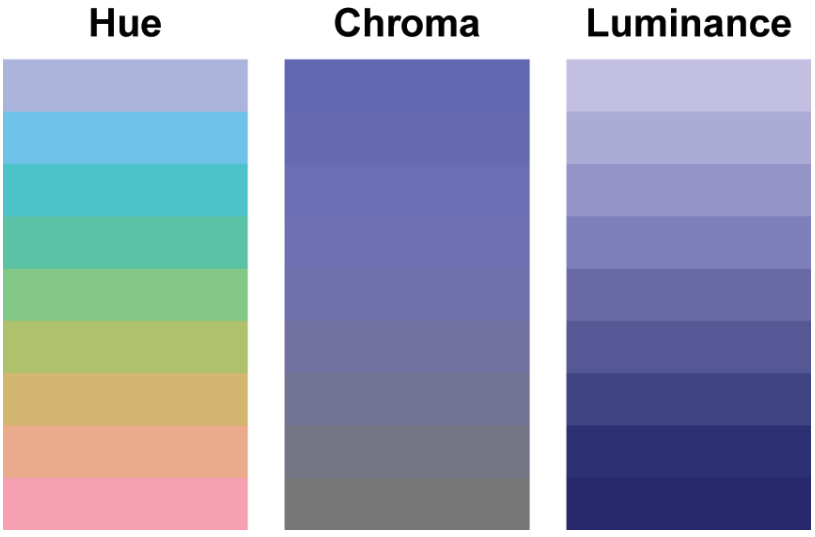

Fig. 1. The three dimensions of the HCL color model: hue, chroma, and luminance. In each panel, one dimension (see heading) changes linearly across the corresponding axis while the others are held constant.

lack of perceptual ordering, it not only confuses the reader but also obscures data through its inability to present small details and might even actively mislead the reader. As an exception from the rule Ware (1988, 49) suggested, "If you wish to read metric quantities using a color key, then a sequence that does not vary monotonically with the color opponent channels should be used. A good example is a spectrum approximation." However, considering the guidelines in the next section and the examples shown in this paper, one can see that for a wide range of purposes a spectral (rainbow) scheme is not the best choice.

GUIDELINES FOR EFFECTIVE COLOR MAPS. In recent years, several publications created guidelines for how to use colors effectively. Although those guidelines differ slightly, there are some cornerstones on how to create effective visualization. Before showing some real-world examples, it is worth introducing these rules (see Ware 2004; Rogowitz and Treinish 1996; Brewer 1997; Rogowitz and Treinish 1998; Treinish 1998; Light and Bartlein 2004; Hagh-Shenas et al. 2007):

Spatial frequency: High-frequency (detailed) data are best represented by monochromatic color maps that only differ in luminance (Mullen 1985).

Form: The human brain is extremely efficient in gathering the shape of an object. This information mainly comes from differences in luminance; therefore, form (e.g., terrain information) will be most effectively coded in the luminance dimension.

Number of colors: For classification tasks (search and distinguishing), only a small number of different hues can be processed with a low error rate. Healey (1996) showed only five to seven different hues can be found accurately and rapidly on a map. Furthermore, MacEachren (1995) wrote that, if the task is to precisely identify a certain color in a plot, the detection rate can plummet when the number of colors increases (detection rate for 10 colors: 98%; for 17 colors: 72%).

Data: Color should be seen more as an attribute of an object than as its primary feature. The human brain is more effective in capturing shape, form, position, lengths, or orientation than in gathering different colors. Therefore, plain plot types should be used if possible (e.g., line, bar, or box plots; see Carswell and Wickens 1990). Additional color can support the reader/analyst if the color matches the data. For continuous variables (e.g., temperature, total number of people in a region), sequential schemes are very effective (Figs. 2b,c). Isoluminant

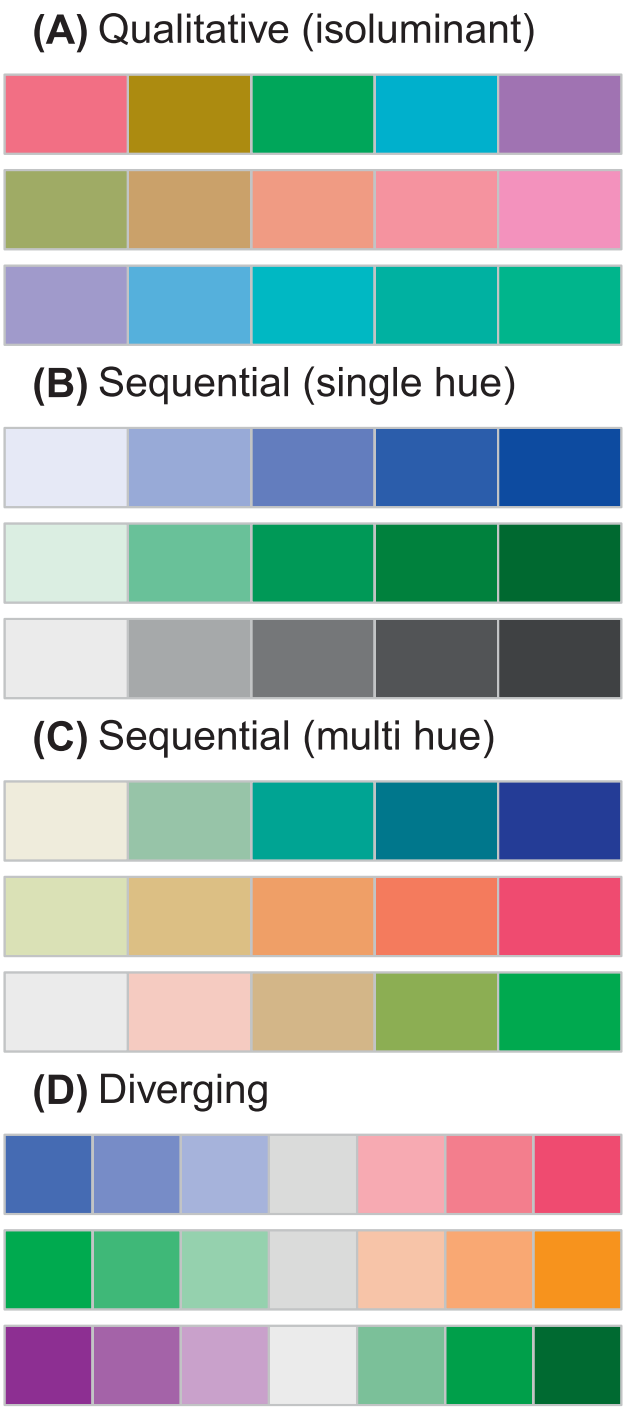

Fig. 2. Examples of different HCL color maps: (a) isoluminant qualitative schemes (e.g., for classification); (b),(c) sequential color maps (e.g., for continuous data: for either increasing or decreasing data with only one extreme); and (d) diverging schemes (e.g., for data with two extremes centered around a neutral value). Sequential schemes can contain one single hue or by passing from one to another trough the HCL color space along the hue dimension. The single-hue sequential color maps shown in (b) are all based on the identical luminance function. Even if they do have different/no hue, the grayscale representations of those three examples are exactly equal. Beside those main types, mixed/hybrid color maps are also possible.

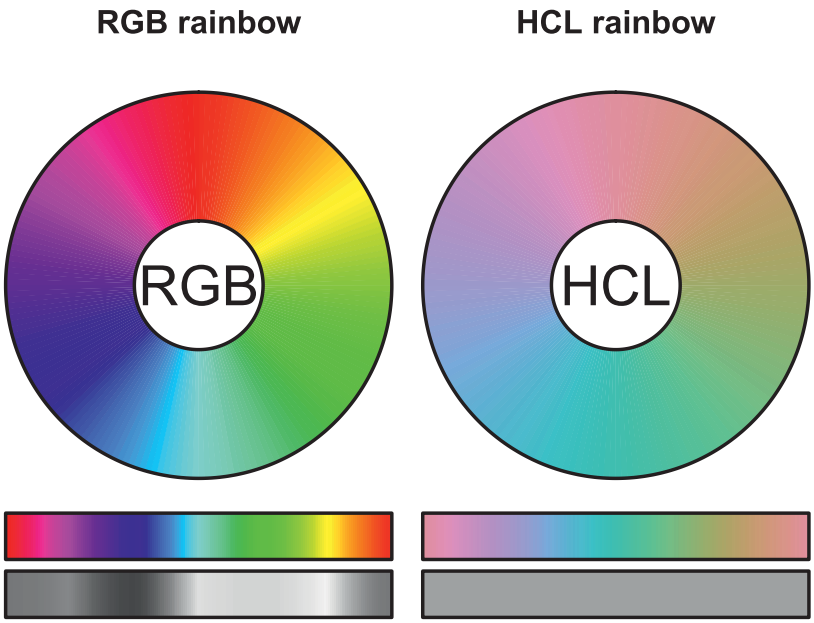

Fig. 3. Juxtaposition of the RGB rainbow color map and an HCL-based rainbow. Below the color wheel, the same palette is shown as a color bar in the colorized version and the corresponding desaturated version, respectively. The RGB rainbow creates unwanted variations in luminance, while the HCL rainbow is fully isoluminant.

qualitative schemes (Fig. 2a) work best for classification because they do not add perceptional distortion to the data. For data with a well-defined neutral value (e.g., precipitation anomalies, balance data), a diverging color scheme with a neutral color around this center point works well (Fig. 2d).

Unique hues: The opponent color theory describes six colors where two colors build an opponent pair at a time. Those pairs are black–white, green–red, and blue–yellow. Our visual system is very efficient at separating opponent colors. If only two different colors are necessary, a pair of opponent colors might be a good choice. For figures that are just containing symbols or markers and just a small set of colors, the six colors of the opponent color theory might work well. However, note that for people with a color deficiency this task can get impossible (cf. "Meteorological examples: Dealing with visual constraints" section).

Contrast: Objects and distinct shapes are easier to identify if there is a clear boundary between them and the surroundings. If necessary, this can be achieved by adding additional contours with a high contrast to the colors of the pattern at the boundary.

Background: The objects (the information) and the background should clearly differ in luminance. Furthermore, the background color should be neutral (white, light gray, or black) not to skew the colors.

Heterogeneity: An object will stand out as a distinct figure if there is a difference in the background.

Conventions: If there are conventions, they should be taken into account (e.g., hot = red, cold = blue, high alert level = red, etc.). However, such conventions are not available for all different purposes and they can strongly differ between cultures.

User task: One of the most important issues is to be aware of the purpose of the user of the visualization: who the end users are (e.g., professional scientists, residents of a country, or decision makers like a civil protection service), what prior knowledge they possess, and what their requirements are. The second driving factor is the type of information that should be transmitted. The most effective color palettes can differ completely—whether we have to communicate thresholds, continuous data, or an abstract or detailed representation of the environment. Some examples are discussed later in the article.

To summarize. As the guidelines show, effective colors have to fulfill a variety of requirements. Although those requirements are rather guidelines than rigorous rules, breaking them can rapidly diminish the effectiveness of the corresponding display. The content of most visualizations is complex enough. Colors should not amplify that. Sometimes it can be a benefit that highly saturated vivid colors shine out; on the other hand, they can produce a lot of "colorjunk" (Tufte 1990). A figure with countless colors and needless luminance gradients makes it harder to gather the important information. The major task of color mapping is to guide the reader and to capture her/his interest. Not losing the reader with unappealing colors is an advantage and should be one of the aims. Furthermore, the task of the user strongly affects the choice of colors, and it is self-evident that figures for a scientific article and a popular product for an Internet platform can have different demands. A final important controlling feature is the medium on which the visualization will be transported (color representation and resolution). In an ideal case, the colors should work

everywhere including screens, data projectors, and printers (grayscale and colorized).

METEOROLOGICAL EXAMPLES. With the conceptual understanding of how the HCL color model works and the guidelines shown above we can now discuss some common meteorological products and plot types to demonstrate the benefits of a more perceptional color concept. The first one is about increasing continuous data: forecasted precipitation amounts. The second example deals with categorical data: a severe weather advisory. The third has multiple contents and is a map used for air mass and frontal analysis.

What can go wrong with inefficient color maps? Let us start with Fig. 4, showing a 5-day accumulated precipitation forecast over the East Coast of the United States during the landfall of Hurricane Sandy in 2012. Figure 4a shows the original figure as provided by the National Oceanic and Atmospheric Administration (NOAA) on its public website.

If one is already familiar with this special type of product, one can quickly identify regions with low precipitation amounts (vivid green colors) and those with high amounts of precipitation (reddish colors), but it is hard to grasp the message of the whole figure at once. For all those not familiar with the colors, it is even harder. One has to scan the entire image and compare the map with the color bar to get an idea of the information shown. The reason, therefore, is the vigorous varying underlying luminance and the high number of different mostly fully saturated colors and very strong color gradients (see box/marker a). Both features actively mislead the reader and make it difficult to capture the entire information. Figure 4b shows the desaturated representation of the original figure. The way luminance changes does not support the visual construction of the data as a form or virtual surface. Furthermore, one can see that the color

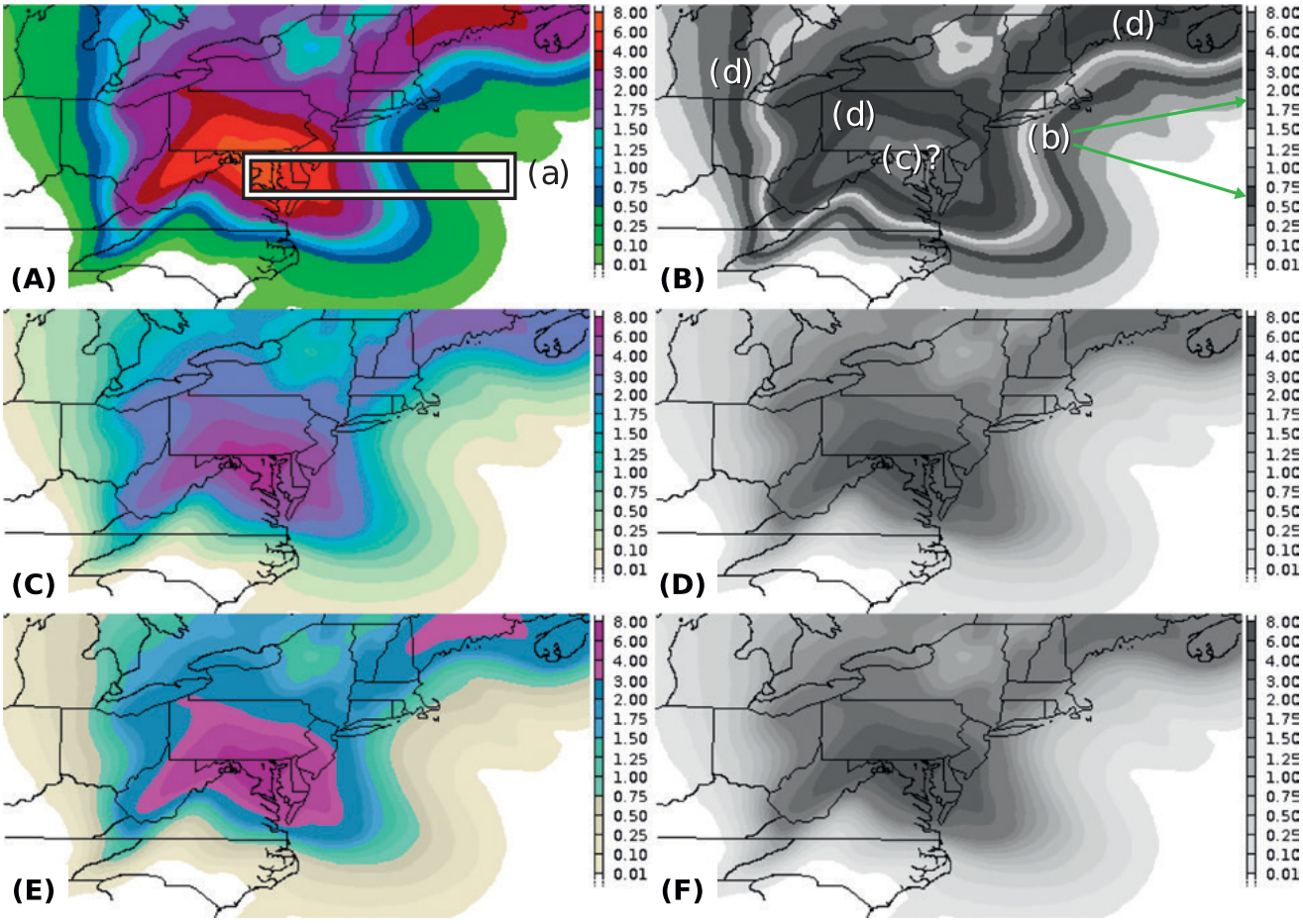

Fig. 4. A rainfall amount forecast during the landfall of Hurricane Sandy on 29 Oct 2012, over the U.S. East Coast. The data are shown in inches accumulated over 120 h: (left) the colorized version and (right) its grayscale representation. (a) The original version as provided by NOAA (www.noaa.gov/). (c),(e) Alternative color maps based on the HCL color concept. The reason for this color choice, the assumed user task, and markers a–d shown are discussed in the paper.

concept completely breaks down when displayed on a monochromatic medium (e.g., grayscale print). The allocation between gray tones and values is no longer unique (marker b in Fig. 4b), the maximum value is no more obvious (marker c), and the overall information gets strongly distorted. Readers are automatically focusing on the distinct dark bands (marker d).

But how can this be improved? Let us have a look at Fig. 4d. This is the desaturated version of an alternative HCL-based color scheme where the values are directly coded in the luminance dimension. Low values have a high luminance (toward white) while luminance monotonically decreases with increasing amounts of precipitation. In contrast to the version above, the human visual system can rapidly recognize the overall shape of the data. Furthermore, the most important parts of the map stand out. This helps the reader/analyst to identify the most important parts as quickly as possible, even without checking the color bar at the outset. The guidance is additionally supported by changing chroma and hue in Fig. 4c. Areas with low values—in this case the less important regions—fade out toward the white background while the regions of interest stand out in luminance, chroma, and hue (dark, colorful, and reddish). As the original figure, the new color map contains 13 different colors but now with a smooth transition from one side to another without creating unwanted distortions. The monotone transition in all three dimensions of the HCL color space creates the impression of a smooth and continuous form or surface. If you compare Figs. 4c and 4e, you can see that we modified how hue and chroma changes over the full palette. The reason is that we redefined the main user task in Fig. 4e. Local communities or civil defense organizations may be interested in some critical thresholds. The guidelines suggest that we can rapidly distinguish small numbers of different hues. Therefore, we combined two color map concepts. While the absolute values are still coded in the luminance dimension (to obtain the shape of the data), hue and chroma are now based on four different categories (gray/green/blue/reddish; stepwise increasing chroma and hue). The result is a hybrid color map as shown in Fig. 4e, a mixture between a qualitative and a sequential scheme. Our visual system can rapidly distinguish the areas indicated by the different hues and chromas, but we are still able to capture the overall shape or form of the data shown or to translate a specific color into its value if necessary. Because we have not changed the way how the luminance changes from low to high values the grayscale representation of both Figs. 4e and 4f is exactly the same. The example shows that the choice of the color map is strongly connected to the user task. However, taking care of some basic guidelines can help to improve the way the information is transported.

Dealing with visual constraints. Color blindness, or color vision deficiency, is another important aspect when choosing effective colors (Brettel et al. 1997; Harrower and Brewer 2003; Light and Bartlein 2004). In Europe, about 8% of the male population has visual constraints (slightly less in the United States; see Miles 1943; Wong 2011; Fairchild 2013b). Far more men than women are affected. Besides the relatively rare monochromacy (light/dark contrasts only), two main types of dichromacy or constrained trichomacy are observed among the male population: Either one of the cone cell subclasses is lacking entirely (about 2%) or it is anomalous (about 6%). The most frequent of these is the deuteranomaly, also known as red–green blindness caused by a cell anomaly. People with this type of anomaly are poor at discriminating small changes in hues in the red–yellow–green spectrum.

Again, we would like to show you a real-world example to illustrate what can happen if visual constraints of the end user are not considered. Figure 5 shows a warning map for Austria in 2013 for severe precipitation amounts. The left column shows the original image while the right column shows an alternative HCL-based color palette, both traversing from green via yellow, orange, and red to purple. The top row shows the colorized version followed by a desaturated version thereof and emulated deuteranope vision (red–green blindness) in the bottom row.

Let us start with the colorized version in Fig. 5a: Like in the example before, all colors are on maximum saturation to attract the attention of the reader. Warning maps, or warning products in general, are often colored similarly to replicate the colors of a traffic light (to regard conventions). The disadvantage of the chosen colors in Fig. 5a: it is hard to capture the most important areas. The vivid colors all over the color map coerce us to scan the whole image. As an alternative, we show an HCL-based color map. The data shown here are a mixture of qualitative data (classification) and continuous data with one extreme (highest alert level). Because only five different colors are necessary a multihue sequential scheme seems to be convenient with a decreasing luminance toward a one-sided extreme at higher alert levels. To take care of the traffic light concept we kept the hues going from green to magenta. To get an effective color map with a steady color change all colors have to lie on a path between the two anchor points

within the HCL color space (light green on one side and dark magenta on the other). Because of the shape of the HCL color space, it is possible that the path lies beyond the boundaries of the HCL gamut. If this happens it is necessary to adjust the dimensions. In our case, luminance and chroma have to be tuned until all colors are well defined in HCL dimensions. Therefore, it is necessary to pick more "pastel" colors, which leads to the color map shown in Fig. 5d. In return, we can strongly improve the reader support and guidance of the product.

In the desaturated versions in Figs. 5b,e, while the HCL version still conveys the essential information, the desaturated RGB version shows something different. Yellow colors are rather light, which results in a higher luminance than the surrounding orange and green (which are barely distinguishable). An inappropriate representation of the warning levels is the result. Similarly with emulated deuteranope vision (Figs. 5c,f), while the interpretation of the original RGB version gets difficult or impossible, the HCL version preserves full readability. This aspect is very important for some fields of application. The example shown here warns the inhabitants of Austria of severe weather situations. Imagine that for 4 out of 100 people it can be very difficult to gather the information just because of inefficient colors. Using more appropriate color palettes can make it much easier for them to interpret the plots and to gain the important information. Clearly, it is important to think about for whom the product should be accessible and to tailor visualizations for the needs of end users.

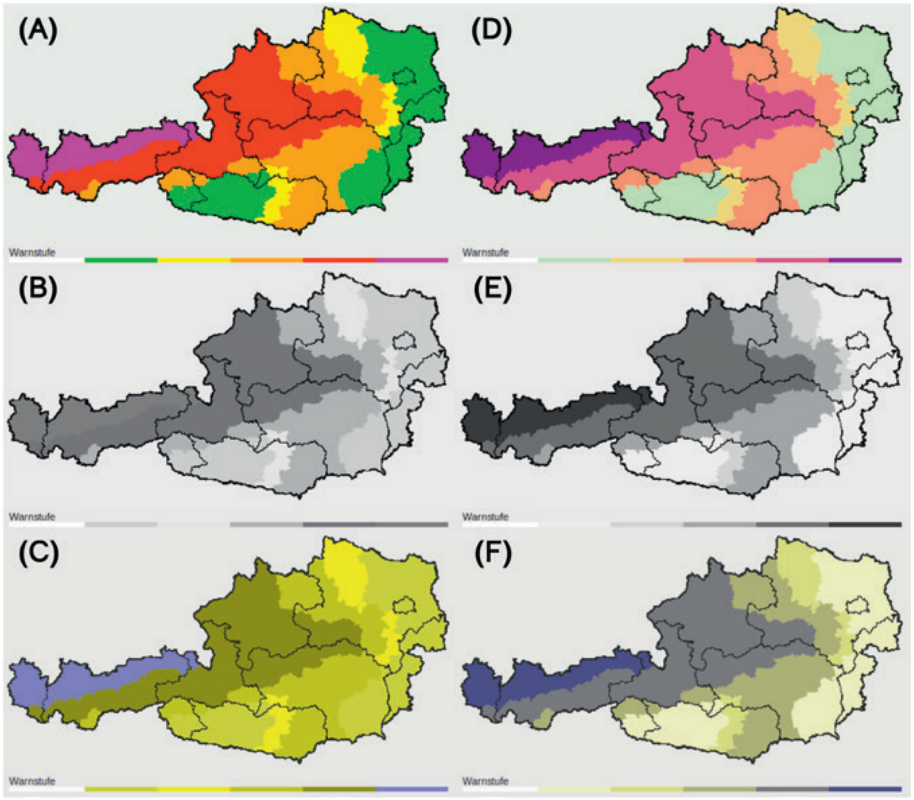

Fig. 5. A severe weather advisory for Austria published on 31 May 2013. (a) The original image as published by UBIMET GmbH (2013), slightly modified because of further postprocessing steps. (d) A modified version with HCL-based colors. (b),(e) As in (a),(d), but desaturated version. (c),(f) Simulation of the appearance for people with deuteranomaly (red–green weakness). Because of the lack of perceptual representation of the RGB color space, different distortions can be found in (a)–(c).

Supporting and guiding the specialized user. To give a broader idea of how to make use of the HCL, we consider a somewhat more complex example where the user task is to identify fronts. Figure 6 shows an equivalent potential temperature analysis on 700 hPa from the European Centre for Medium-Range Weather Forecasts (ECMWF) as employed in the internal weather platform at the Institute of Meteorology and Geophysics in Innsbruck. Equivalent potential temperature is a conserved variable and widely used to identify different air masses and fronts. Fronts are found at regions of large gradients of equivalent potential temperature. Additionally, geopotential height is shown to appraise the movement of the air to identify the front types (Steinacker 1992).

Figure 6a shows the product as it was provided over the last decade using a rainbow-type color map as found in many other meteorological websites and products. Because of the strong color gradients, especially between red and green (opponent colors), a large proportion of our less experienced students was misled and placed the fronts at color boundaries instead of at the strongest equivalent potential temperature gradients. Mostly, fronts were allocated to the areas where red and green encounter

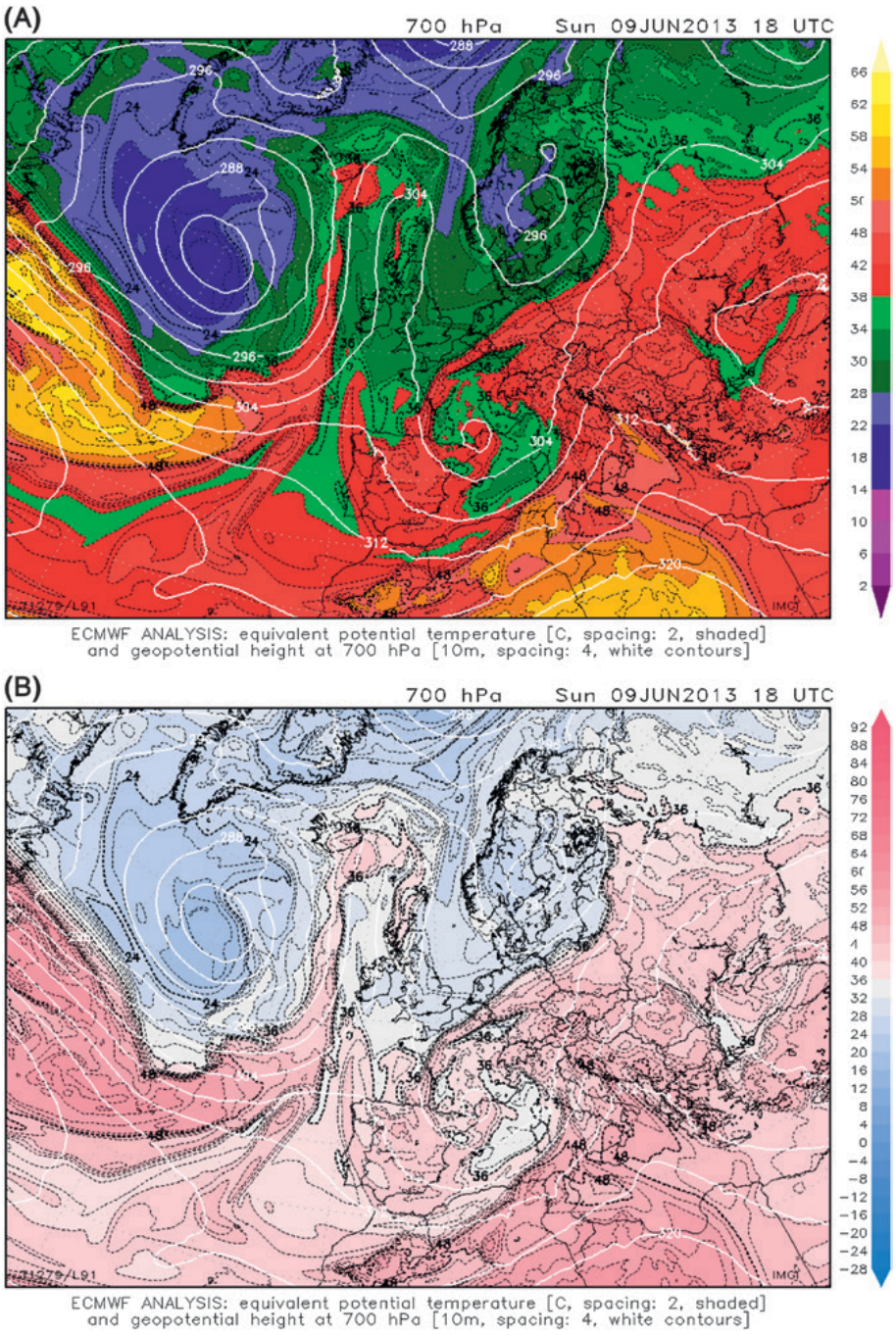

Fig. 6. ECMWF analysis of the equivalent potential temperature (°C) at 700 hPa over the Atlantic/Europe. (a) The old product based on highly saturated RGB rainbow colors. (b) The revamped product including an HCL-based color map.

each other because the color gradients obscured the physical gradients. The striking quantity of misinterpretations was the main motivation to redesign our products.

Figure 6b shows the new appearance of the same analysis field. Please remember that the main user task is to identify frontal zones. The exact absolute (metric) values of the equivalent potential temperature are not of interest. To get an overall idea of the air masses (cold/warm), we added a diverging color scheme using the conventional colors (red/blue) centered around the empirical model climate mean for the displayed area. The hue in Fig. 6b has to be seen as an additional attribute supporting the reader to capture the overall distribution of air masses and to identify the two extremes as quickly as possible. Additionally, three types of contour lines are shown. A high contrast (black) was used to code the important information: borders for the geographical orientation and the isentropes for the frontal analysis. As secondary information, geopotential height is shown in white contours. Because this information is less important on the first look, it is forced into the background.

The original RGB-based product is a good example how colors can actively mislead the reader. Since we removed this shortcoming, the number of misinterpretations decreased by roughly 50% (empirical value from the daily weather-briefing lecture at the Institute of Meteorology and Geophysics, University of Innsbruck; G. J. Mayr 2013, personal communication).

TOOLS AND FURTHER READING. We have selected three cases from among hundreds of meteorological products to demonstrate how to apply coloring guidelines and how to make use of color palettes derived from the HCL color space. Because a good concept is worth little without ease of use, we now make some suggestions of how to integrate the HCL color concept into your daily workflow. Even if visualization software does not provide the HCL scheme, it should be emphasized that each HCL

color can be converted to the corresponding RGB coordinates with the corresponding hexadecimal representation. The authors use the colorspace package (Ihaka et al. 2013) written in R, but there are packages and sources for other established coding languages as well.

Online tool. We set up an online interface to create customized palettes. The tool "online HCL creator" is available online (www.hclwizard.org/). The interface offers some typical examples of statistical maps/graphics and some specifically meteorological chart types. You can easily modify the color palettes and tune them for your personal needs. Furthermore, the tool gives you the ability to emulate the appearance of the chosen colors under different visual constraints (e.g., for deuteranope viewers) or in a desaturated representation (e.g., on a grayscale printer). The interactive examples give a first impression of what the resulting color maps look like. Moreover, we developed some export functions for common software languages so that the HCL palettes can be comfortably applied to your own data in a familiar software environment.

Advanced users. One of the most powerful tools to create HCL palettes for all possible uses is the package colorspace (Ihaka et al. 2013), which provides various types of color space manipulations/transformations plus an intuitive graphical user interface (GUI) to pick color maps. The colorspace package is written in R (R Core Team 2013), an open-source programming language that has also been receiving increasing attention within our community. Our online interface mentioned above is based on this colorspace package, mimicking its choose_palette() graphical user interface. There are also modules for other languages such as Python (see colormath module; Taylor 2014) or MATLAB (see image processing toolbox or the colorspace transformation module; Getreuer 2011) that allow you to choose and transform color sequences in different color spaces.

Other tools. Harrower and Brewer (2003, 2011) developed the online tool ColorBrewer.org that provides predefined color palettes for various purposes with focus on map makers. While their palettes are not directly based on the HCL model, the guidelines used in their creation are very similar resulting in often similar sets of colors. Easy-to-use sets of colors are available in different coding languages (e.g., colorbrewer in Python, RColorBrewer in R, cbrewer in MATLAB). However, the disadvantage of ColorBrewer.org is that you can only pick from preset color schemes.

CONCLUSIONS. Visualizations are used regularly to communicate methods, data, and findings. The ubiquitous availability of computers and software enables everyone to create all possible types of visualizations and animations, yet creating effective plots and maps is not a trivial task. When colors are used, they are mostly derived from the famous red–green–blue (RGB) color space because most software offers easy access to RGB-based palettes and RGB-based color map designers. Consequently, a large proportion of the science community uses RGB-based color palettes uncritically (Borland and Taylor 2007; Light and Bartlein 2004; Rogowitz and Treinish 1998).

In this article, we offer basic guidelines on how to use color more effectively in visualizations. Therefore, we introduce the less well known hue–chroma–luminance (HCL) color concept as a toolbox to achieve this goal. In contrast to the technical RGB color model, the HCL color concept is based on how human color vision works (Zeileis et al. 2009; Silva et al. 2011). The HCL color model captures the three main perceptional dimensions—hue (dominant wavelength: defining the color), chroma (colorfulness: compared to gray), and luminance (brightness: amount of gray). With these three dimensions, a broad range of colors can be defined (Knoblauch 2002; Fairchild 2013a), facilitating specification of effective colors for various purposes.

To demonstrate the advantages of the HCL over the RGB color space we discussed three common visualization types from the meteorological field and demonstrated that the HCL color model can help to define effective color maps for all kinds of visualizations. The benefits are as follows: better readability; full functionality in grayscale/luminance; enhanced end user support; more effective conveyance of complex concepts; and enhanced accessibility for people with visual constraints. But any color model fails when the user task is not considered. It is probably the most crucial factor dictating how to choose the most effective colors for a given product. If the user task is unknown or not well defined, the effectiveness of a figure can get lost completely.

Additionally, the presented tools should help to easily adapt the proposed concepts for your own work. In particular, we provide a web interface to the R package colorspace where everyone can create personal HCL color maps and export them in different formats for several common software languages.

We have often experienced skepticism about changing familiar color maps when introducing potential users to the HCL colors. However, much more often than not, this skepticism turned into enthusiasm after a few days of using HCL-based products, especially with the availability of tools to ease implementation into one's own workflow.

ACKNOWLEDGMENTS. This study was supported by the Federal Ministry for Transport, Innovation and Technology (BMVIT) and Austrian Science Fund (FWF): TRP 290-N26. The first author was also supported by a Ph.D. scholarship from the University of Innsbruck, Vizerektorat für Forschung.

REFERENCES

BAMS, 1993: Guidelines for using color to depict meteorological information: IIPS subcommittee for color guidelines. *Bull. Amer. Meteor. Soc.,* **74,** 1709–1713, doi:10.1175/1520-0477(1993)0742.0.CO;2.

Borland, D., and R. Taylor, 2007: Rainbow color map (still) considered harmful. *IEEE Comput. Graphics Appl.,* **27,** 14–17, doi:10.1109/MCG.2007.323435.

Brettel, H., F. Viénot, and J. D. Mollon, 1997: Computerized simulation of color appearance for dichromats. *J. Opt. Soc. Amer.,* **14A,** 2647–2655, doi:10.1364/JOSAA.14.002647.

Brewer, C. A., 1997: Spectral schemes: Controversial color use on maps. *Cartogr. Geogr. Inf. Syst.,* **24,** 203–220, doi:10.1559/152304097782439231.

Carswell, C., and C. Wickens, 1990: The perceptual interaction of graphical attributes: Configurality, stimulus homogeneity, and object integration. *Percept. Psychophys.,* **47,** 157–168, doi:10.3758/BF03205980.

Fairchild, M. D., 2013a: Color appearance terminology. *Color Appearance Models,* John Wiley & Sons, 85–96.

——, 2013b: Human color vision. *Color Appearance Models,* John Wiley & Sons, 1–37.

Getreuer, P., cited 2011: Colorspace transformations. [Available online at www.mathworks.com/matlabcentral/fileexchange/28790-colorspace-transformations.]

Hagh-Shenas, H., S. Kim, V. Interrante, and C. Healey, 2007: Weaving versus blending: A quantitative assessment of the information carrying capacities of two alternative methods for conveying multivariate data with color. *IEEE Trans. Visualization Comput. Graphics,* **13,** 1270–1277.

Harrower, M. A., and C. A. Brewer, 2003: ColorBrewer.org: An online tool for selecting color schemes for maps. *Cartogr. J.,* **40,** 27–37, doi:10.1179/000870403235002042.

——, and ——, 2011: ColorBrewer.org: An online tool for selecting colour schemes for maps. *The Map Reader: Theories of Mapping Practice and Cartographic Representation,* M. Dodge, R. Kitchin, and C. Perkins, Eds., John Wiley & Sons, 261–268.

Healey, C., 1996: Choosing effective colours for data visualization. *Proc. Visualization,* San Francisco, CA, IEEE, 263–270.

Ihaka, R., P. Murrell, K. Hornik, J. C. Fisher, and A. Zeileis, cited 2013: Colorspace: Color space manipulation. [Available online at http://CRAN.R-project.org/package=colorspace.]

Knoblauch, K., 2002: Color vision. *Steven's Handbook of Experimental Psychology—Sensation and Perception,* S. Yantis and H. Pashler, Eds., Vol. 1, 3rd ed. John Wiley & Sons, 41–75.

Light, A., and P. J. Bartlein, 2004: The end of the rainbow? Color schemes for improved data graphics. *Eos, Trans. Amer. Geophys. Union,* **85,** 385–391, doi:10.1029/2004EO400002.

MacEachren, A., 1995: How maps are seen. *How Maps Work: Representation, Visualization, and Design,* Guilford Press, 51–147.

Mahy, M., L. Van Eycken, and A. Oosterlinck, 1994: Evaluation of uniform color spaces developed after the adoption of CIELAB and CIELUV. *Color Res. Appl.,* **19,** 105–121.

Miles, W. R., 1943: Color blindness in eleven thousand museum visitors. *Yale J. Biol. Med.,* **16,** 59–76.

Mullen, K. T., 1985: The contrast sensitivity of human colour vision to red-green and blue-yellow chromatic gratings. *J. Physiol.,* **359,** 381–400.

R Core Team, 2013: R: A language and environment for statistical computing. R Foundation for Statistical Computing Rep., 3604 pp. [Available online at http://web.mit.edu/r_v3.0.1/fullrefman.pdf.]

Rogowitz, B., and L. Treinish, 1996: How not to lie with visualization. *Comput. Phys.,* **10,** 268–273, doi:10.1063/1.4822401.

——, and ——, 1998: Data visualization: The end of the rainbow. *IEEE Spectrum,* **35,** 52–59, doi:10.1109/6.736450.

Silva, S., B. Sousa Santos, and J. Madeira, 2011: Using color in visualization: A survey. *Comput. Graphics,* **35,** 320–333, doi:10.1016/j.cag.2010.11.015.

Smith, A. R., 1978: Color gamut transform pairs. *Proc. Fifth Annual Conf. on Computer Graphics and Interactive Techniques,* Atlanta, GA, SIGGRAPH, 12–19.

Steinacker, R. A., 1992: Dynamical aspects of frontal analysis. *Meteor. Atmos. Phys.,* **48,** 93–103, doi:10.1007/BF01029561.

Stevens, S., 1966: Matching functions between loudness and ten other continua. *Percept. Psychophys.,* **1,** 5–8, doi:10.3758/BF03207813.

Taylor, G., cited 2014: Python-colormath. [Available online at http://python-colormath.readthedocs.org.]
Treinish, L. A., 1998: Task-specific visualization design: A case study in operational weather forecasting. *Proc. Visualization,* Research Triangle Park, NC, IEEE, 405–409.
Tufte, E., 1990: *Envisioning Information.* Graphics Press, 126 pp.
UBIMET GmbH, cited 2013: Österreich–Alle Warnungen. [Available online at www.uwz.at/at/de/karte/alle_warnungen.]
Ware, C., 1988: Color sequences for univariate maps: Theory, experiments and principles. *IEEE Comput. Graphics Appl.,* **8,** 41–49, doi:10.1109/38.7760.
——, 2004: Color. *Information Visualization: Perception for Design,* Morgan Kaufmann, 103–149.
Wong, B., 2011: Color blindness. *Nat. Methods,* **8,** 441, doi:10.1038/nmeth.1618.
Zeileis, A., K. Hornik, and P. Murrell, 2009: Escaping RGBland: Selecting colors for statistical graphics. *Comput. Stat. Data Anal.,* **53,** 3259–3270, doi:10.1016/j.csda.2008.11.033.

Article II

Stauffer, R., Mayr, G.J., Messner, J.W., Umlauf, N. and Zeileis, A. (2017). *Spatio-Temporal Precipitation Climatology over Complex Terrain Using a Censored Additive Regression Model*. International Journal of Climatology, 37, 3264–3275, doi:10.1002/joc.4913.

JCR ranking: **Category 1** in *Meteorology & Atmospheric Sciences*.

INTERNATIONAL JOURNAL OF CLIMATOLOGY
Int. J. Climatol. **37**: 3264–3275 (2017)
Published online 9 November 2016 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/joc.4913

RMetS
Royal Meteorological Society

Spatio-temporal precipitation climatology over complex terrain using a censored additive regression model

Reto Stauffer,[a,b]* Georg J. Mayr,[b] Jakob W. Messner,[a,b] Nikolaus Umlauf[a] and Achim Zeileis[a]
[a] *Department of Statistics, Faculty of Economics and Statistics, University of Innsbruck, Austria*
[b] *Institute of Atmospheric and Cryospheric Sciences, University of Innsbruck, Austria*

ABSTRACT: Flexible spatio-temporal models are widely used to create reliable and accurate estimates for precipitation climatologies. Most models are based on square root transformed monthly or annual means, where a normal distribution seems to be appropriate. This assumption becomes invalid on a daily time scale as the observations involve large fractions of zero observations and are limited to non-negative values.

We develop a novel spatio-temporal model to estimate the full climatological distribution of precipitation on a daily time scale over complex terrain using a left-censored normal distribution. The results demonstrate that the new method is able to account for the non-normal distribution and the large fraction of zero observations. The new climatology provides the full climatological distribution on a very high spatial and temporal resolution, and is competitive with, or even outperforms existing methods, even for arbitrary locations.

KEY WORDS climatology; precipitation; complex terrain; GAMLSS; censoring; daily resolution

Received 4 April 2016; Revised 8 July 2016; Accepted 12 September 2016

1. Introduction

Accurate knowledge of precipitation climatology is important for a wide range of applications, such as agriculture, risk assessments, strategic project planning, water resource management, and tourism. Moreover, climatological information is often used as background information for statistical downscaling, or as a baseline for model verification. For locations equipped with a precipitation measurement instrument, this task is straightforward. However, the observational network is generally too sparse to capture all local effects, and observations are preferentially located at lower elevations and close to populated areas due to environmental conditions and maintenance purposes.

To gain information about the amount or occurrence of precipitation for locations without measurements, information from an irregularly spaced observation network has to be brought to a finer (regular) region-wide grid trough interpolation. Thiessen (1911) pointed out that simple interpolation schemes, such as nearest neighbour, or arithmetic areal means, should not be used for interpolation of precipitation as these methods do not account for local factors that affect precipitation, e.g. distance to mountain ranges, geographical position, and others (Basist *et al*., 1994). Thiessen (1911) invented an areal weighted-mean scheme that includes terrain-based properties. Although this was proposed as a first 'simple' extension, today's statistical methods still follow a similar idea.

Over recent decades, several different approaches have been developed, which can be clustered into three main classes. The first class consists of 'exact interpolation schemes', including inverse distance weighting, and various forms of Kriging (e.g. Biau *et al*., 1999; Goovaerts, 2000). Inverse distance weighting is often not suitable as dependencies on topography, for example, cannot be considered. For Kriging several extensions exist to include additional covariates, or spatio-temporal Kriging (Snepvangers *et al*., 2003; Aryaputera *et al*., 2015). The second class comprises 'regional regression models', where for every location a (simple) regression model is adjusted from only a subset of neighbouring stations. Examples are PRISM (Precipitation-elevation Regressions on Independent Slopes Model; Daly *et al*., 1994, 1997, 2002, 2008), and Daymet (Thornton *et al*., 1997).

A third class of interpolation methods consists of 'smooth spline regression models', which are the focus of this article. Generalized additive models (GAMs; Guisan *et al*., 2002) are a common form of smooth spline models, where a response quantity is described by a set of possibly nonlinear functions of covariates. Feasible functions could be an altitudinal effect, a cyclic effect to represent the seasonality, or a two-dimensional spline on longitude and latitude to describe the spatial distribution. Spline models have already been used for long-term climatologies for different quantities, such as temperature or precipitation (e.g. Boer *et al*., 2001; Jarvis and Stuart, 2001; Vicente-Serrano *et al.*, 2003; Guan *et al*., 2009).

However, regarding precipitation, most studies have focused only on monthly or even annual means.

* Correspondence to: R. Stauffer, Department of Statistics & Institute of Atmospheric and Cryospheric Sciences, University of Innsbruck, Universitätsstrasse 15, A–6020 Innsbruck, Austria. E-mail: reto.stauffer@uibk.ac.at

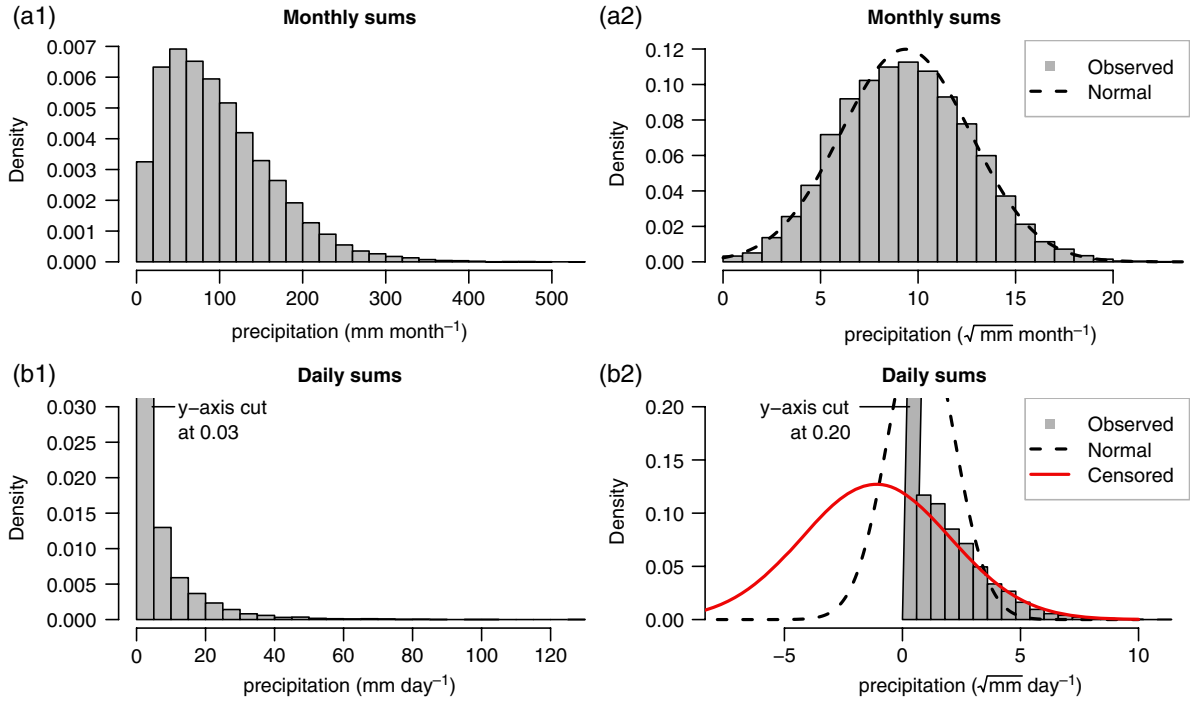

Figure 1. Density plot of precipitation sums. Top row: monthly precipitation sums from all 117 stations. Bottom row: daily precipitation sums of one sample station ('Iselsberg-Penzelberg'). Right column: power-transformed observations ($\sqrt{\text{mm}}$ day^{-1}) with a fitted normal distribution (black, dashed). In addition, a fitted zero left-censored normal distribution is shown for the daily power-transformed observations (b2; red, solid). Please note: *y*-axes in the bottom row are both cut. [Colour figure can be viewed at wileyonlinelibrary.com].

Finer than monthly temporal resolution is needed for a wide range of applications, so climatologies limited to monthly sums are unsatisfying. Furthermore, additional useful properties of the climatological distribution are of great interest, such as the probability of precipitation, or specific quantiles. This can be achieved by either creating a specific statistical model for each of the quantity of interest, or by modelling the full climatological distribution in one model. Such fully distributional climatological estimates can be used for statistical downscaling approaches based on, for example, quantile mapping or model output statistics. Quantile mapping is often used to calibrate climate or weather forecast models by re-sampling the forecasted distribution from the observed distribution (Themeßl *et al*., 2012; Acharya *et al*., 2013; Ajaaj *et al*., 2016; Rajczak *et al*., 2016). Climatological estimates are also used as background information for spatial model output statistics methods to account for site-specific climatological features not yet resolved by the numerical weather or climate models (Scheuerer and Büermann, 2014; Dabernig *et al*., 2016; Stauffer *et al*., 2016). Furthermore, climatological estimates are useful as baseline verification. Fully probabilistic spatial climatologies provide all necessary information to compute a wide range of verification scores, such as Brier scores, root mean squared errors (RMSE), but also probabilistic scores like the continuous rank probability score (CRPS), or quantile score.

An accurate estimate of the full climatological distribution requires a suitable response distribution. Figure 1(a1) shows an example of monthly precipitation sums of 117 stations, which are strongly skewed to the right (positive skewness). To remove the skewness, a power transformation has been used frequently (Box and Cox, 1964). In literature, cubic (Stidd, 1973) or square root (Hutchinson, 1998a) transformations have often been suggested but may vary for different climatic zones or temporal aggregation periods. After applying a square root transformation (Figure 1(a2)) the majority of the skewness is removed and the data are close to a normal distribution (dashed line). Aggregation into monthly sums usually moves the data away from zero yielding a pseudounbounded data set (Sansom and Tait, 2004), wherefore the assumption of a normal distribution might be appropriate. However, daily precipitation sums show different properties. Figure 1(b1) shows all daily sums of station 'Iselsberg-Penzelberg' (later referred to as station B). Three main properties can be identified:

i The distribution is strongly positively skewed,
ii the distribution is limited to non-negative values (≥0), and
iii a large fraction of all observations is exactly zero (dry days).

To remove the characteristic skewness a power transformation can be applied, the remaining properties must be accounted for separately by assuming a proper response distribution. For hourly or daily precipitation sums that remain physically limited to non-negative values after the power transformation (Figure 1(b2)). Several studies have shown that the concept of censoring works well for precipitation, as precipitation is physically limited

to non-negative values (Messner *et al*., 2014; Scheuerer, 2014; Scheuerer and Hamill, 2015).

In this article, we present a novel spatio-temporal additive model with a zero left-censored normal response, to estimate a full-distributional climatology of precipitation over complex terrain on a daily temporal resolution. To obtain both, the climatological mean and the climatological variance, a distributional regression model (Klein *et al*., 2015) is used. Distribution regression allows all parameters of a response distribution to be modelled by a set of explanatory covariates. Statistical frameworks that allow for distributional regression are often termed as vector generalized additive models (VGAM; Yee, 2015) or generalized additive model for location, scale, and shape (GAMLSS; Rigby and Stasinopoulos, 2005). We use a GAMLSS model to obtain the climatological estimates for each day of the year, and for any arbitrary location within the study area. A power transformation is used to deal with the skewness, while assuming a zero left-censored normal distribution handles both the lower limit at zero, and the large fraction of zero observations in the data set. This new approach allows full scalability (size of the area of interest, but also spatial- and temporal resolution) and can therefore be implemented easily and applied to new data sets or regions.

2. Methodology

2.1. Left-censored normal distribution

The response distribution of the model is crucial to its overall performance. In contrast to the observed monthly sums, even after the power transform, daily values show a strong peak at zero caused by the large fraction zero observations (days without precipitation). The concept of censoring is that the response itself is limited to a certain threshold τ, or a range, and cannot be observed outside these limits. It is assumed that there is a latent unobservable process driving the response, which can be described by suitable covariates. As precipitation is physically limited to 0 mm, it can be seen as left-censored at zero ($\tau = 0$). The resulting zero left-censored normal distribution is specified as follows:

$$y = \max\ (0, y^*), \quad y^*\ \sim N(\mu, \sigma) \tag{1}$$

y^* denotes the unobservable 'latent' response following a normal distribution, given the parameters location μ and scale σ. The 'observable' response y is simply the maximum of the latent response and the censoring point. From here on this distribution will be denoted as N_0. The density (φ_{cens}) and the distribution function (Φ_{cens}) for N_0 can be written as follows:

$$\varphi_{\text{cens}}\left(x_i|\mu,\sigma,0\right) = \begin{cases} 0 & \text{for } x_i < 0 \\ \Phi\left(x_i|\mu,\sigma\right) & \text{for } x_i = 0 \\ \varphi\left(x_i|\mu,\sigma\right) & \text{else} \end{cases} \tag{2}$$

$$\Phi_{\text{cens}}\left(x_i|\mu,\sigma,0\right) = \begin{cases} 0 & \text{for all}: \quad x_i < 0 \\ \Phi\left(x_i|\mu,\sigma\right) & \text{else} \end{cases} \tag{3}$$

While both quantities are set to zero below the censoring point, both follow the density Φ and distribution function Φ of a non-censored normal distribution, respectively, above the censoring point ($x_i > 0$). On the censoring point ($x_i = 0$), the distribution function is again equivalent to the normal distribution, while the density represents the probability that an observation will lie exactly on zero. Therefore, the probability π to exceed zero can be written as:

$$\pi\,(y > 0) = 1 - \Phi\,(0|\mu,\sigma) \tag{4}$$

A last property of interest is the expectation of N_0. As the estimates will be fitted on a power-transformed scale y with $y = z^{1/p}$, this transformation has to be included to get the expectation on the original scale (mm day^{-1}). The expectation function of a power-transformed N_0 can be expressed as (Appendix):

$$E\,[z] = \int_0^\infty z \cdot \varphi\left(z^{\frac{1}{p}}|\mu,\sigma\right) \cdot \frac{z^{\left(\frac{1}{p}-1\right)}}{p} \mathrm{d}z \tag{5}$$

where z is the observable response on the original scale, μ and σ are the estimated parameters of N_0 on the power-transformed scale, and p denotes the power parameter of the power transformation.

Figure 1(b2) shows the fitted zero left-censored normal distribution (solid line) which is able to depict the distribution of the data very accurately. Comparing the estimated (44%) and observed (43%) probability of precipitation, as well as the estimated (2.3 mm day^{-1}) and observed (2.2 mm day^{-1}) expectation shows that the power-transformed zero left-censored normal distribution is able to account for the large fraction of zero observations, and to accurately adjust the distribution of the non-censored part.

2.2. Generalized additive models for location, shape, and scale

GAMLSS (Rigby and Stasinopoulos, 2005) are an extension to GAMs (Guisan *et al.*, 2002) that allow all parameters of a certain response distribution to be modelled separately. In case of a censored normal distribution two parameters have to be specified: 'latent' location (mean), and 'latent' scale (standard deviation). For a zero left-censored normal distribution (N_0) the GAMLSS model can be expressed as follows:

$$\begin{aligned} y &\sim N_0\,(\mu,\sigma) \\ \mu &= s\,(\mathbf{x}) \\ log\,(\sigma) &= t\,(\mathbf{x}) \end{aligned} \tag{6}$$

The 'observable' response y is assumed to follow N_0 with location μ and scale σ, where the log-link ensures positive scale values during optimisation. Both parameters can be expressed by a set of unknown, possibly nonlinear

functions $s(\dots)$ and $t(\dots)$, also known as linear predictors. The explanatory variables $\mathbf{x}$ include the covariates, such as altitude, longitude, latitude, or others.

In the GAMLSS framework, the linear predictors can include different additive effects, such as linear effects, nonlinear effects, cyclic effects, or two-dimensional surfaces. For nonlinear one-dimensional and multi-dimensional effects, splines are used very frequently. Common forms of splines are, e.g. thin-plate splines, or B-splines (Wood, 2006, Chap. 4.1, Fahrmeir *et al*., 2013). As complex splines tend to get wiggly, an additional penalization term is estimated yielding to smooth regression splines.

For applications where only the mean is of interest, the scale parameter in Equation (6) could be specified as a constant. The result would be a homoscedastic GAM model where the variance is constant among all observations. Models of this type have been used frequently for the application of precipitation climatologies, such as in Hutchinson (1998a, 1998b); Price *et al.* (2000); Boer *et al.* (2001); Hong *et al.* (2005). However, as we would like to estimate the full daily climatological distribution, the linear predictor for $log(\sigma)$ in Equation (6) has to be specified in addition.

For the specific application of a spatio-temporal precipitation climatology, the effects $s(\mathbf{x})$ and $t(\mathbf{x})$ have to capture a possible altitudinal effect, the seasonality, as well as the spatial pattern. Therefore, the following effects have been specified for parameter μ (location):

$$\mu = s(\mathbf{x}) = \beta + s_1(\text{alt}) + s_2(\text{yday}) + s_3(\text{lon}, \text{lat}) + s_4(\text{yday}, \text{lon}, \text{lat}) \quad (7)$$

where β denotes the global intercept, s_1(alt) represents a smooth 'altitudinal' effect, s_2(yday) a cyclic seasonal effect based on the 'day of the year', s_3(lon, lat) a two-dimensional spatial effect given the geographical coordinates 'longitude' and 'latitude', and s_4(yday, lon, lat) represents a three-dimensional spline to account for spatial variabilities of the seasonal pattern across the region of interest. Cubic splines are used for the seasonal effect as they allow addition of a cyclic constraint such that no discontinuities emerge between December and January. All other effects use thin-plate regression splines. Thin-plate regression splines use the Eigenbasis of the data and 'provide optimal low rank approximations to thin-plate splines that are both computationally efficient and stable' (Wood, 2003; p. 110) typically leading to better results.

Analogously to the linear predictor for the location μ (Equation (7)), the linear predictor for the log-scale is expressed as follows:

$$log(\sigma) = t(\mathbf{z}) = \gamma + t_1(\text{alt}) + t_2(\text{yday}) + t_3(\text{lon}, \text{lat}) + t_4(\text{yday}, \text{lon}, \text{lat}) \quad (8)$$

The two linear predictors include the same effects, as we expect the climatological variance for precipitation to also show a seasonal and spatial dependency, as well as an altitudinal effect (Equations (7) and (8)). This seems appropriate for the specific task of this article, but is no general requirement for GAMLSS models.

2.3. Model setup

To estimate the nonparametric smooth model as specified in Equations (6)–(8) suitable software is required which allows for a zero left-censored normal distribution. We are using a novel *R* package 'bamlss' (Umlauf *et al*., 2016b) which offers a flexible Bayesian framework for additive models for location, scale, and shape (and beyond), and the capability to handle (very) large data sets. Other possible software implementations to estimate smooth models are, e.g. ANUSPLIN (Hutchinson, 2014), or the R packages 'mgcv' (Wood, 2006) and 'gamlss' (Rigby and Stasinopoulos, 2005).

In addition to the full spatio-temporal climatology station-wise models are estimated for comparison. These station-wise models use the same technique, but as only the information of one station is used, the spatial effects are not required. The station-wise models therefore only include the intercepts (β, γ) and the seasonal effects [s_2(yday), t_2(yday)], while all other assumptions are equivalent to Equations (6)–(8).

The skewness is removed by applying a power transformation to the observations. In previous studies, the empirical power parameters $p = 2$ (square) or $p = 3$ (cubic) have been used. However, the power parameter depends on the data and the response distribution of the application. To obtain the best power parameter for this study, we fitted one station-wise GAMLSS model for each station in the data set, optimizing the regression coefficients plus an additional power parameter simultaneously. Optimal power parameter did not show an obvious spatial or altitudinal dependency and varied between $p = 1.3$ and $p = 2.0$. Tests have shown that the model performance is not very sensitive within this range, therefore, a fixed value of $p = 1.6$ corresponding to the median of all estimated power coefficients was chosen.

The estimates for all GAMLSS models ('station-wise' and 'spatio-temporal'; Section 4) are based on the new *R* package bamlss. The optimisation is based on Markov-Chain Monte Carlo (MCMC) sampling in combination with an iterative weighted least squares backfitting algorithm (Umlauf *et al*., 2016a). Code and data used in this article can be downloaded from the bamlss project page (http://bayesr.r-forge.r-project.org/).

3. Area of interest and data

This article focuses on the temperate alpine state of Tyrol, Austria, located in Central Europe. Tyrol lies in the Eastern Alps and consists of two separated parts – North Tyrol located north of the main Alpine ridge, and East Tyrol located south of the main Alpine ridge, as shown in Figure 2. The topography reaches from 465 up to 3798 m amsl including the majority of the highest mountains in Austria. This complexity is one of the main

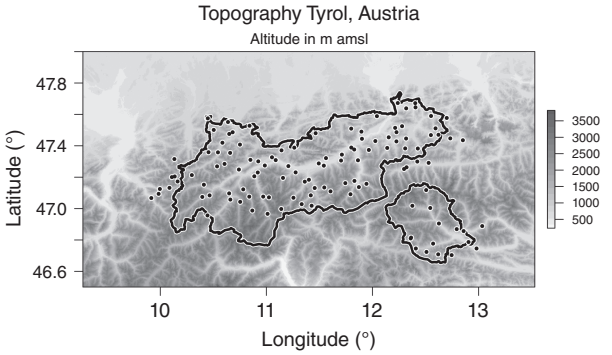

Figure 2. Topography around Tyrol, Austria. Shading indicates altitude of the topography in metres, the outline shows the border of the state of Tyrol consisting of North and East Tyrol. The black dots show the stations locations from the data set.

difficulties from a climatological perspective, as climatological properties can strongly vary within just a few kilometres due to topographically induced effects.

Compared to other regions, Tyrol has a relatively dense precipitation observation network with a mean station distance of about 10 km. The observation data set used in this study is provided by the local hydrographical service and includes 117 stations (Figure 2) spanning September 1971 through the end of 2012. A total of 78 out of 117 stations include at least 40 years of data, 14 start within the 1980s, nine within the 1990s and three post-millennial. Each station is equipped with a manual rain gauge to measure liquid water or liquid water equivalent accumulated over the last 24 h, observed at 0600 UTC. The hydrographical service performs rigorous quality controls on the observations. The total data availability is about 88% equating to *ca* 1.6 million unique daily observations. For a region with such a complex topography, a dense observational network is essential to be able to depict all small-scale spatial features, and also the pronounced altitudinal effects. For regions with less complex topographic features, a sparser network might be sufficient as the spatial differences within the area will presumably be smaller. The data set used in this study is freely available for non-commercial use, and can be downloaded from the bamlss project page (BMLFUW; http://bayesr.r-forge.r-project.org/).

Figure 3 shows the mean monthly precipitation sums for all stations. The largest amounts of precipitation with around 1100–2100 mm per year are observed for the north-west and north-east stations, and a second slightly weaker maximum with >1000 mm per year for the south-east stations. This is due to the proximity to the foreland of the Alps (Bavaria, Germany to the north, northern Italy to the south) and dynamically driven processes. Incoming air masses are lifted when they encounter the first obstacles, leading to orographic precipitation, and a loss of moisture at the foot of the Alps (Houze, 2012). On the north side, this effect is mainly caused by fronts advected from north-westerly directions, leading to higher mean precipitation amounts over the whole year. In the south-east, the highest precipitation amounts are related to mesoscale cyclones forming over the Mediterranean sea (e.g. Raulin, 1879; Frei and Schär, 1998). All stations show a local maximum in summer (June–August), which is mainly caused by local thermal convection, which leads to increased amounts of precipitation and thunderstorms. The convective enhancement is strongest in the pre-alpine regions north-west, and north-east of Tyrol (Wapler, 2013).

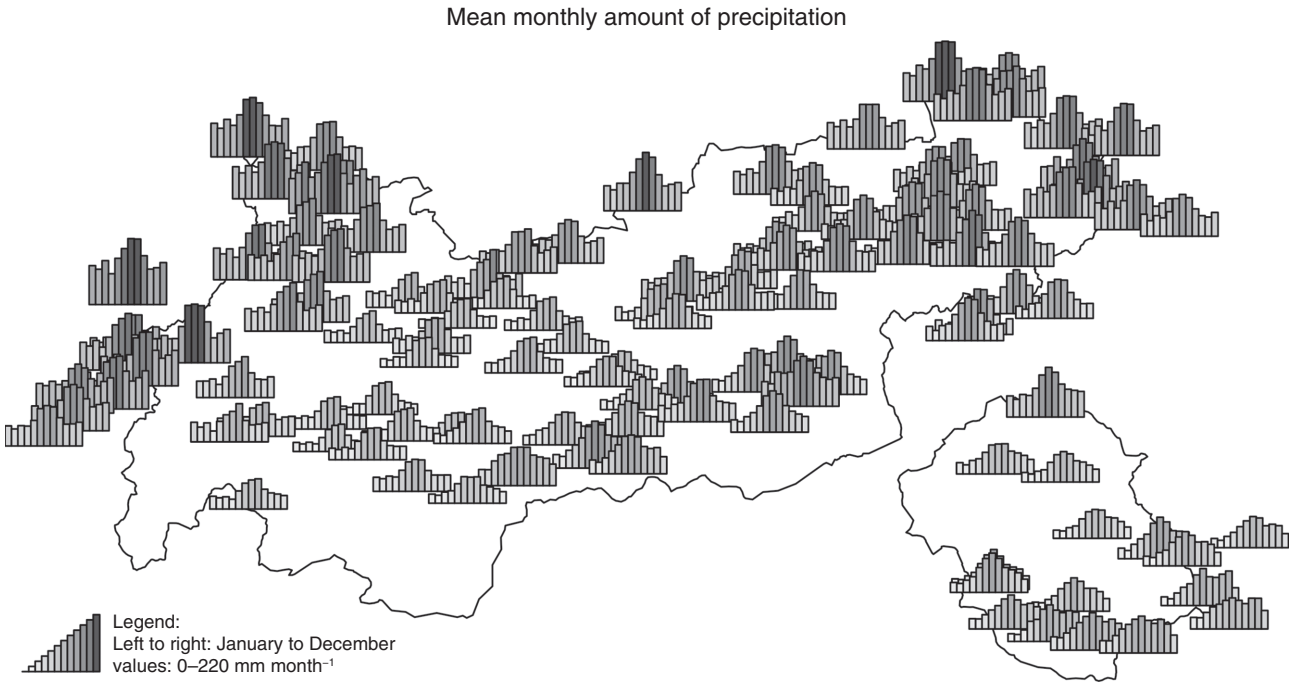

Figure 3. Mean monthly precipitation based on the data set. Each bar indicates 1 month (January–December, left to right). Bar height and luminance contain the same information (0–220 mm month^{-1}).

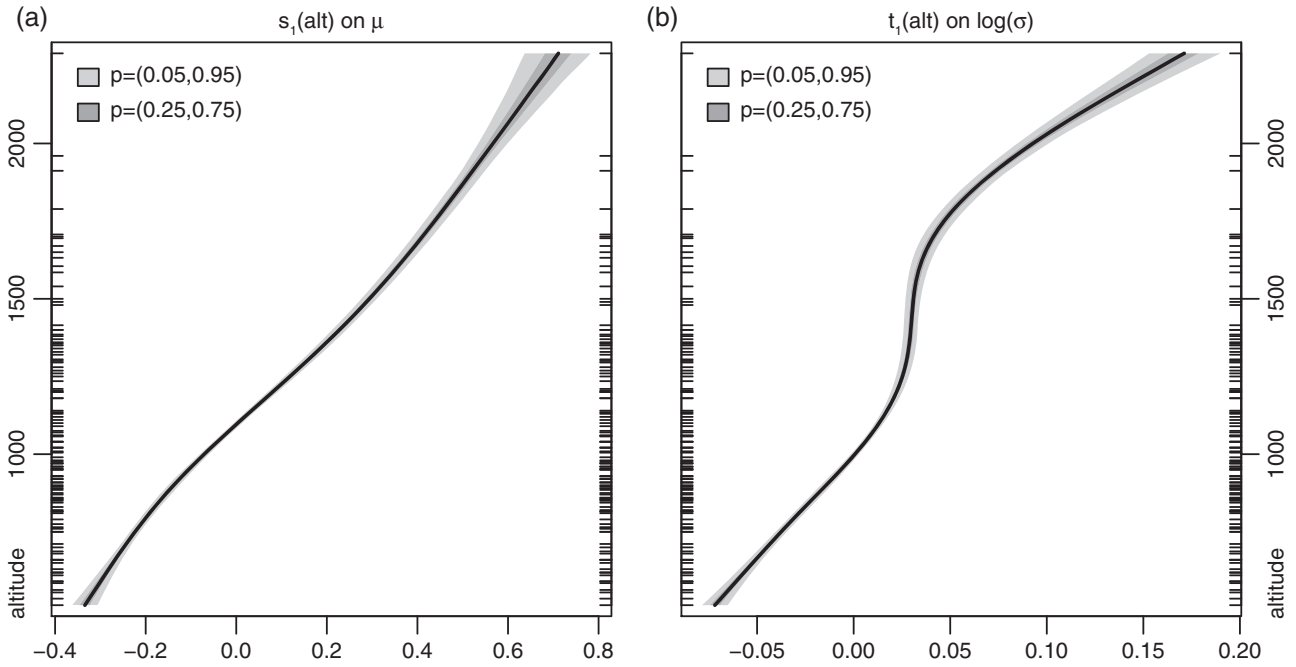

Figure 4. Centred altitudinal effects $s_1(\text{yday})$ on location μ (a), and $t_1(\text{yday})$ on $log(\sigma)$ (b). Values on the power-transformed scale. Inner ticks on the ordinate indicate the altitudes of all stations in the data set. The shading shows the confidence intervals of the estimate and the width is closely related to the large amount of training data.

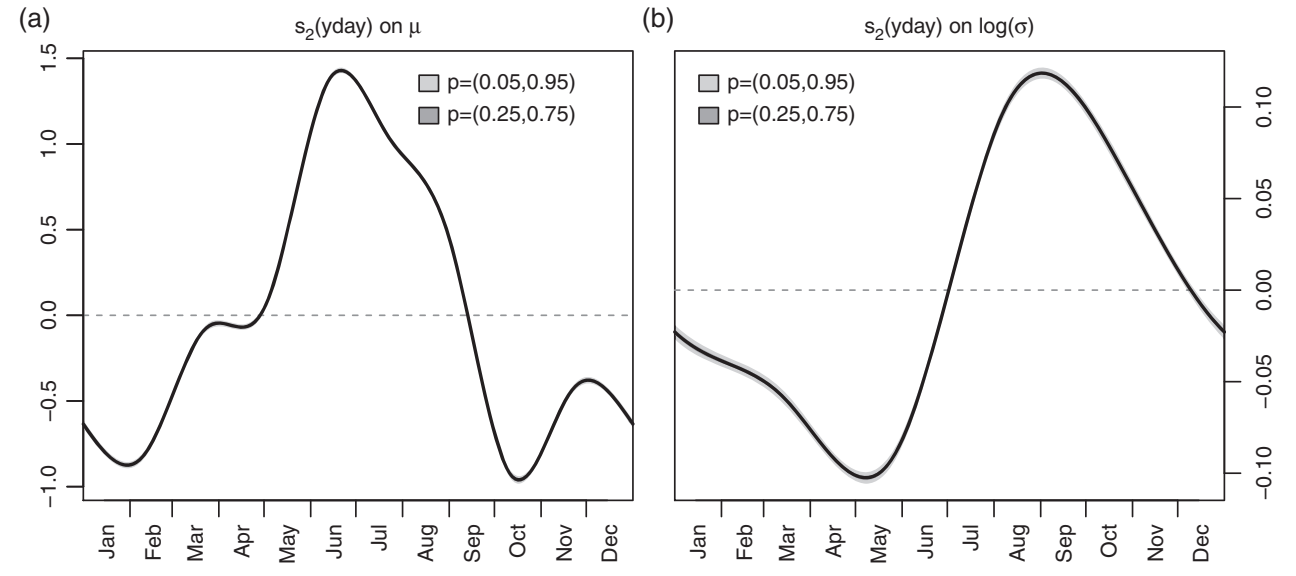

Figure 5. Centred cyclic seasonal effects $s_2(\text{yday})$ on location μ (a), and $t_2(\text{yday})$ on $log(\sigma)$ (b). Values on the power-transformed scale. The shading shows the confidence intervals of the estimate, the width is closely related to the large amount of training data. The effect controls the global seasonal effect for all stations.

4. Results

First, the estimated effects of the new censored spatio-temporal precipitation climatology will be shown in Section 4.1, followed by a model comparison and validation.

4.1. Results of the new daily-based spatio-temporal model

As described in Section 2.3 a spatio-temporal GAMLSS with a zero left-censored normal response is used to create the long-term precipitation climatologies (Equation (6)) with the linear predictors for location (μ) and log-scale ($log(\sigma)$) as specified in Equations (7) and (8). The individual effects of the two linear predictors are shown in Figures 4–7. All figures, except the last, show centred effects on the power-transformed scale.

Figure 4 shows the altitudinal effects for location μ (left), and log-scale (right). As expected, the amount of precipitation and the variance increase with increasing altitude (Ekhart, 1948; Frei and Schär, 1998). The global

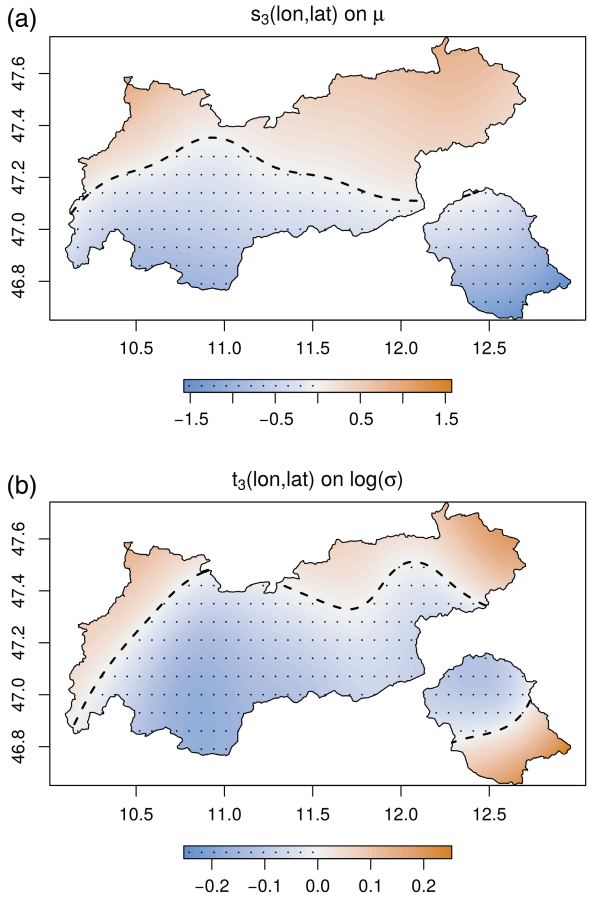

Figure 6. Centred spatial effect $s_2(\text{lon}, \text{lat})$ on location μ (a), and $t_3(\text{lon}, \text{lat})$ on $log(\sigma)$ (b). Values on the power-transformed scale. Positive values orange, negative values blue and additionally dotted. The effect controls the mean underlying climatological spatial distribution of precipitation. [Colour figure can be viewed at wileyonlinelibrary.com].

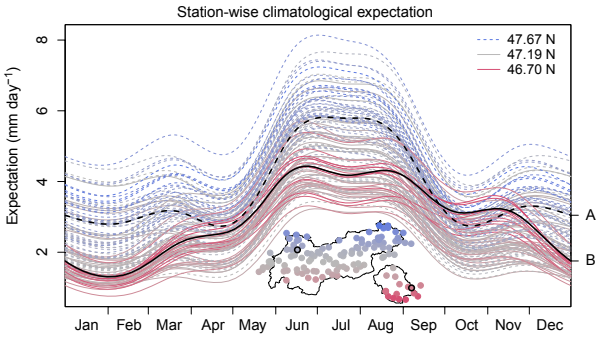

Figure 7. Expectation in mm day^{-1} for all 117 stations used. Stations are coded blue/dashed to the north, and red/solid to the south. The two sample stations A and B as shown in Figure 8 are highlighted in black. The difference in the seasonal pattern between north and south results from the tri-variate thin-plate splines s_4, t_4 based on the day of the year, longitude, and latitude. [Colour figure can be viewed at wileyonlinelibrary.com].

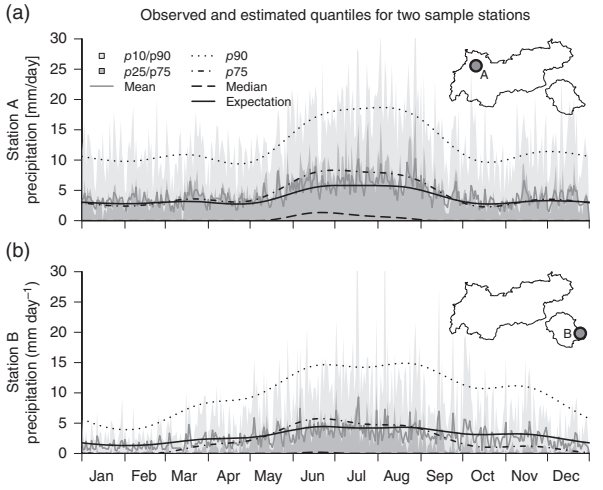

Figure 8. Distribution of daily observed precipitation sums for two sample stations. The long-term daily distribution is shown in grey including 42 years of observations: 10–90% and 25–75% inner-quantile ranges (shaded), and mean (solid, grey). In addition, the climatological estimate from the spatio-temporal model is shown. Expectation (Equation (5)) as solid, and quantiles as black lines of different styles. Mean annual precipitation sums/frequency of observed precipitation for both stations: station A 'Namlos' (a) 1577 mm year^{-1}/48%, station B 'Iselsberg-Penzelberg' (b) 954 mm year^{-1}/36%. Table 1 contains CP for both stations.

cyclic seasonal effects for location μ (left) and log-scale (right) are shown in Figure 5. The seasonal effect shows the overall dry winter conditions from December to February (compare Figure 3) with a low variability. Overall, June–August are the months showing largest amounts of precipitation, with increasing variability during mid to late summer related to the convective season with its peak between July and September. During this time period, location μ already decreases, while the log-scale nearly reaches its overall maximum. Or in other words: in autumn, the overall amount of precipitation strongly decreases (relatively dry), but the variability reaches its local annual maximum. October is the driest month but still shows high variability compared to the first half of the year.

The spatial effects are shown in Figure 6. As for the seasonal cycle, location μ and log-scale show different patterns. While location μ increases from south to north, the log-scale effect reaches its maximum towards the pre-alpine plains with Bavaria, Germany to the north, and Italy to the south. The increase in location μ is related to fronts reaching Tyrol predominantly from north and north-westerly directions. The increase in the variability is mainly caused by higher convective activity (Wapler, 2013), and the orographic precipitation produced when air masses approaching from plains encounter the first higher obstacles.

Seasonal patterns differ for different regions. The three-dimensional thin-plate splines s_4, t_4 in Equations (7) and (8) allow for a spatial variation of the cyclic seasonal pattern across the area of interest. This effect can be seen in Figure 7, which shows the estimated climatological expectation in mm day^{-1} for all 117 stations in the data set. The results show that the new climatology is able to capture the different seasonal characteristics between the sub-regions north and south of the main Alpine ridge.

As the new climatology returns estimates for the full distribution, it is also possible to examine other properties, such as quantiles or the probability of precipitation:

Table 1. CP for station A and station B (Figure 8), and for all 117 stations (overall) for the intervals [0.10–0.90], [0.25–0.75], and [0.00–0.50]. Sample CPs for the spatio-temporal GAMLSS (top) and the station-wise GAMLSS (bottom). Theoretical CPs are shown in parenthesis.

Interval	Station A	Station B	Overall
Spatio-temporal GAMLSS			
0.10–0.90 (0.80)	0.78	0.82	0.81
0.25–0.75 (0.50)	0.46	0.51	0.50
0.00–0.50 (0.50)	0.49	0.51	0.49
Station-wise GAMLSS			
0.10–0.90 (0.80)	0.80	0.81	0.81
0.25–0.75 (0.50)	0.49	0.50	0.50
0.00–0.50 (0.50)	0.49	0.49	0.49

Figure 8 shows the climatological distribution and the corresponding climatological estimates for two sample stations of the data set. Station A is located north of the main Alpine ridge and close to the pre-alpine foreland. Station B lies south of the main Alpine ridge. A few distinct features can be identified. Station A receives precipitation more frequently and observes larger amounts of precipitation than station B. Furthermore, the different seasonality can be seen. While station A shows a clear summer-signal with a strong increase during May–June and a corresponding decrease in autumn, station B shows a smoother transition across the year, with an overall lower amplitude. The censored daily spatio-temporal climatology captures the main features of amplitude, seasonality, and the overall distribution.

In addition, Table 1 contains coverage probabilities (CP) of the spatio-temporal estimates for the two sample stations in Figure 8. The CP shows the fraction of observations falling into the verification interval and therefore shows a measure of calibration. For the two intervals [0.10–0.90] and [0.25–0.75], the theoretical or perfect coverage would be 0.80 and 0.50, respectively. The validation shows 0.78/0.46 for station A and 0.82/0.51 for station B indicating that the estimates are a bit overdispersive (too wide) for station A, and slightly underdispersive (too narrow) for station B.

Figure 9 shows the spatio-temporal climatology for two sample days, 1 January (top), and the 1 June (bottom). The climatological expectation (left column) shows the overall drier winter conditions and the distinct altitudinal dependence with up to $\sim$7 mm day^{-1} on 1 January, and up to $\sim$10 mm day^{-1} on 1 June. The right column shows the probability of precipitation in percent. On 1 January, the highest probability of observing precipitation is towards the foreland to the north, while the inner-alpine regions close to the main Alpine ridge show relatively low probabilities. On 1 June, the overall probability of precipitation increases, with probabilities above $\sim$55% for all mountainous areas.

4.2. Model comparison and validation

The novel spatio-temporal precipitation climatology validated in consideration of a number of aspects of model performance. Of special interest is the performance for fully out-of-sample events to show the predictive performance for future (temporally out-of-sample) events at arbitrary locations within the area of interest (spatially out-of-sample). Consequently, the out-of-sample predictions of the spatio-temporal model will be compared against two in-sample station-wise reference methods. All three models are trained on observations through the end of 2009, including up to 39 years of data (Section 3), evaluated on the remaining 3 years between 2010 and 2012, from here on referred to as 'training' and 'test data set'.

4.2.1. *Monthly mean model*

As a robust and simple baseline reference model, long-term monthly means of the measurements are computed for each station separately. Similarly, the probability of precipitation is the long-term mean frequency of observations greater than zero for a given station and month. Months with missing data are excluded.

4.2.2. *Station-wise GAMLSS*

To validate the goodness of fit of the spatial effects of the spatio-temporal model, station-wise GAMLSS climatologies with a zero left-censored normal distribution have been estimated. One model is estimated for each of the 117 stations using Equations (6)–(8) with modified linear predictors. As these models are station-wise, only the intercepts and seasonal effects have to be included.

4.2.3. *Spatio-temporal GAMLSS*

To score the predictive skill of the novel spatio-temporal climatology, a ten-fold cross-validation is performed. For each cross-fold, a random subset including 10% of all stations is omitted. The spatio-temporal model is estimated on the remaining stations using the specifications of Equations (6)–(8). For the left-out 10% of the stations, the predictions are made on the test data set. This leads to 'spatially out-of-sample' predictions, while both station-wise methods are 'spatially in-sample'.

4.2.4. *Measure of performance*

As a measure of performance, mean absolute errors (MAEs), RMSE, and Brier scores (Brier, 1950) will be shown. While the first two are used for the amount of precipitation, the Brier scores show the performance on the estimated probability of precipitation. MAEs are based on the median of the climatological distribution ($\max(0, y)^p$; Equation (1)), while the RMSE are based on the expectation (Equation (5)). The Brier scores depend on the probability that precipitation will be observed (Equation (4)). A Brier score of zero would indicate a perfect forecast. To compare the different models, error-differences are shown in Figure 10. Each box-whisker is based on 117 values, each of which is the mean error difference of a specific station. The error-differences are shown between each pair of methods,

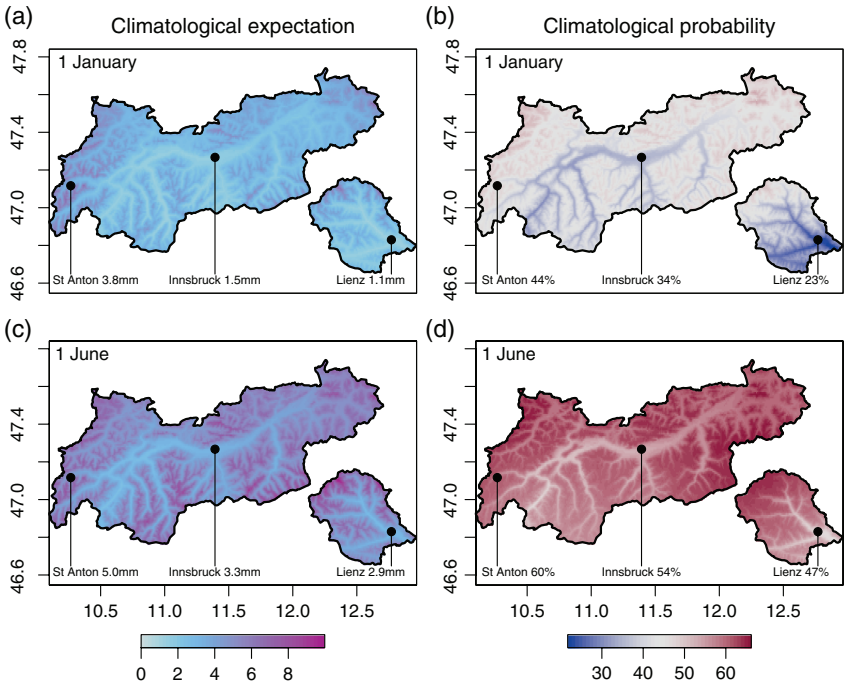

Figure 9. Climatological expectation (a and c; mm day^{-1}), and climatological probability of precipitation (b and d; %) for 1 January (a and b), and 1 June (c and d), respectively. Values are explicitly shown for three locations: St Anton (1284 m amsl), Innsbruck (574 m amsl), and Lienz (673 m amsl). Prediction based on the SRTM DEM (CGIAR-CSI, 2016).

where the difference is defined as 'method B – method A' threading 'method B' as the reference. For example: Figure 10(a) shows the differences in the MAE, where the first pair shows 'monthly mean model (monmean) *versus* station-wise GAMLSS (station)'. On the test data set the 'monthly mean model' performs slightly better than the 'station-wise GAMLSS', while both are more or less identical (in median) evaluated on the training data set. Figure 10(b) and (c) shows the same validation for the RMSE, and Brier score respectively. The novel spatio-temporal zero left-censored GAMLSS model shows comparable results in all measures, or indeed slightly better in terms of Brier scores, even if the predictions of the spatio-temporal GAMLSS model are the only ones which are fully out-of-sample.

A probability integral transform (PIT; Gneiting *et al.*, 2007) histogram is shown in Figure 11 to check the suitability of the fitted climatological distributions. The PIT histogram contains CP for a set of evenly distributed non-overlapping intervals. For each observation, the corresponding quantile of the climatological distribution is evaluated and then pooled into bins. A perfectly calibrated model would show a uniform distribution across all bins. The PIT histogram indicates that the zero left-censored normal distribution seems suitable for the application of precipitation, but the deviation from a perfectly uniform distribution indicates that there is still some room for improvement.

To sum up: the predictive skill of the novel spatio-temporal censored GAMLSS model is competitive in comparison with station-wise estimates, even for regions without observational sites.

The results show that the new spatio-temporal censored GAMLSS model allows to accurate reproduction of the climatology over complex terrain, even for regions without observational sites.

5. Conclusion and discussion

A new method for estimating a spatio-temporal precipitation climatology with a full-distributional response and a daily temporal resolution is presented in this article. The climatology is represented by a GAMLSS using the new *R* package bamlss (Umlauf *et al.*, 2016b) to optimize the regression coefficients. The estimated effects are shown in Section 4.1 and return interpretable and highly significant climatological features. An advantage of a full-distributional model is that a variety of properties can be derived from the estimate. The novel climatology shows a good overall performance for the amount of precipitation on the daily scale, as well as for the probability of precipitation. The results demonstrate that the concept of censoring is suitable to account for the high number of zero observations in the data set. In contrast to the station-wise reference methods shown in the article, the spatio-temporal model returns fully distributional estimates for the whole area, even for regions where no observations are available. The cross-validation shows that the spatial model returns accurate estimates for unobserved regions and is even able to out-perform the station-wise estimates in some cases.

The PIT histogram and the CP show that the model is overall well calibrated, but there is some room for improvement. Further adjustments of all tuning parameters (location μ, scale σ, but also the power parameter p)

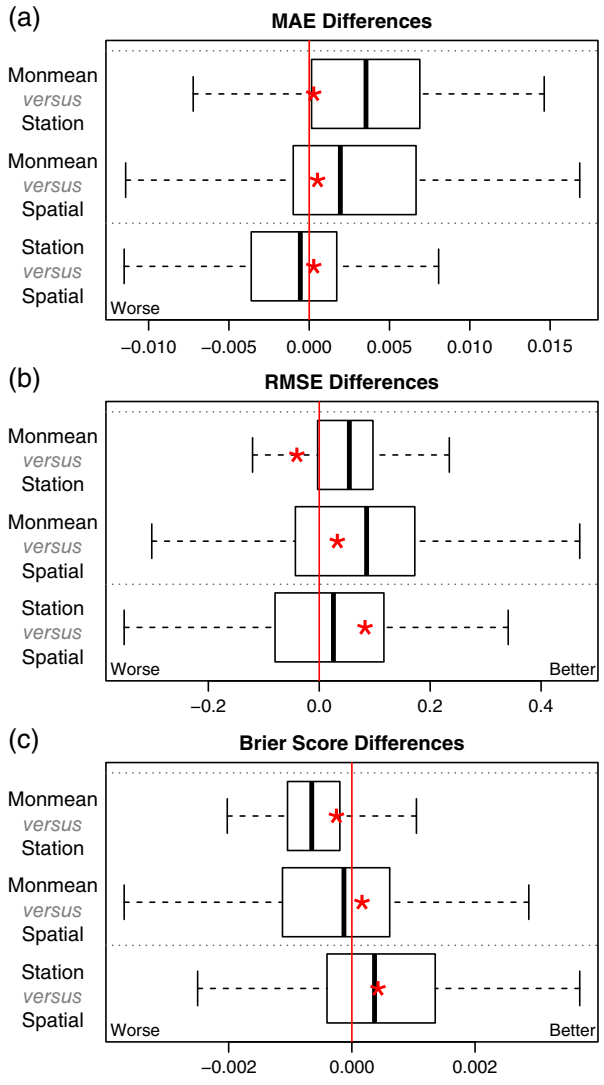

Figure 10. Differences in MAE, RMSE, and Brier scores for all model pairs: monthly mean model (monmean), station-wise GAMLSS (station), and spatio-temporal GAMLSS (spatial). Each box-whisker consists of 117 station-wise values, each of which is the mean error for one specific station. Box-whiskers show the results on the test data set (0.25/0.5/0.75 quantiles plus additional 1.5 inner-quantile range) and the red asterisk indicates the median of the same analysis on the training data set. The differences are defined as 'method B – method A' such that positive values indicate that 'method A' performs better than 'method B' for each 'A *versus* B'. Absolute values lie around 3.35 (MAE), 7.25 (RMSE), and 0.24 (Brier score). [Colour figure can be viewed at wileyonlinelibrary.com].

might have a positive effect on the results. Beside optimizing the parameters of the zero left-censored normal distribution a different response distribution might bring additional benefits. Such distributions could be, e.g. a left-censored logistic distribution (Messner *et al*., 2014), a gamma distribution (Rust *et al*., 2013; Wong *et al*., 2014), or a mixed distribution (Eden *et al*., 2014). Rust *et al.* (2013) has shown that a gamma distribution works well for precipitation on the original scale without the need to apply a power transformation, which might distort the data. On the other hand, the gamma distribution is not defined at zero. While Scheuerer and Hamill (2015) use a censored shifted gamma distribution, Rust *et al.*

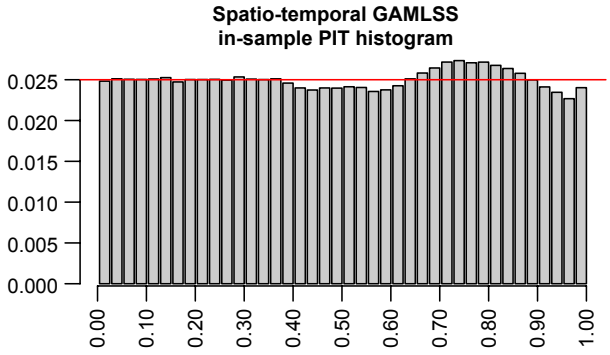

Figure 11. Pit histogram of the spatio-temporal GAMLSS model evaluated on the training data set. Width of the bins: 2.5%. [Colour figure can be viewed at wileyonlinelibrary.com].

(2013) uses a two-part approach where the probability of precipitation is modelled independently from the amount of precipitation. This allows use of the gamma distribution, but has the necessity to define and estimate two different models, while the approach presented in this article requires only the specification of one single model to obtain the full distribution and all its properties.

A direct comparison against more complex existing methods would be needed to explicitly highlight advantages and drawbacks of our method, but needs some extensions to our current model. Adding additional covariates beside the day of the year, longitude, latitude, and altitude could further improve the model results as shown in previous publications. Conceivable covariates could be, e.g. steepness and facing of the slopes, or the distance to the closest open water source. Furthermore, the new model allows inclusion of daily covariates, such as mean wind direction, covariates explaining the regional weather situation, and many others, which is not possible for longer aggregation periods (e.g. monthly). Some covariates have been tested but have not brought the expected results yet. The estimate of the statistical model (Equations (6)–(8)) can be performed in under 32 h on a single core (2.7 GHz Intel Xenon, 8 GB memory), although the model is already quite complex. However, estimating the full model including a 'random' set of covariates will be unsatisfying. One idea would be an automated iterative variable selection approach, such as boosting or ridge-regression, to find the best additional covariates.

One big advantage of the method is that it is fully scalable as the model specifications are very general. While demonstrated for daily precipitation sums in this study, other aggregation periods would also be possible. The model estimation only requires the observations and corresponding covariates. To retrieve full-probabilistic spatial estimates, a suitable digital elevation model (DEM) is needed. The resolution of the DEM predetermines the resolution at which the climatological estimates can be provided, wherefore the results can be served at a very high spatial resolution. As only few inputs are necessary (observations, DEM), and due to the very general framework, the method can easily be applied to other regions and data sets. Additionally, it would be worthwhile to apply the approach

to other censored variables, such as wind speed, sunshine duration, or relative humidity, which would only require minor modifications.

Acknowledgements

Ongoing project is funded by the Austrian Science Fund (FWF): TRP 290. The computational results presented have been achieved in part using the Vienna Scientific Cluster (VSC). Data set is provided by the Federal Ministry of Agriculture, Forestry, Environment and Water Management (BMLFUW), Abteilung IV/4 – Wasserhaushalt (http://ehyd.gv.at).

Appendix

Derivation of the expectation function

Derivation of the expectation function for a power-transformed zero left-censored normal distribution as in Equation (5).

Assume a left-censored normal distribution with a censoring point at zero without a power transformation. The distribution function $F(x)$, and the density function $f(x)$ are defined as follows:

$$F(x),\ f(x) = \frac{\partial F(x)}{\partial x} \tag{A1}$$

and therefore the expectation of the distribution becomes:

$$E[x] = \int_{x=0}^{\infty} x \cdot f(x)\ dx \tag{A2}$$

For a left-censored normal power-transformed distribution, the distribution function $G(z)$ and density function $g(z)$ can be written in the same way, where x from Equation (A1) is simply $z^{1/p}$:

$$g(z) = \frac{\partial G(z)}{\partial z} = \frac{\partial F\left(z^{1/p}\right)}{\partial z} \tag{A3}$$

and therefore:

$$\frac{\partial F\left(z^{1/p}\right)}{\partial z} = f\left(z^{1/p}\right) \cdot \frac{z^{\left(\frac{1}{p}-1\right)}}{p} \tag{A4}$$

leading to Equation (5) as shown in the article.

References

Acharya N, Chattopadhyay S, Mohanty UC, Dash SK, Sahoo LN. 2013. On the bias correction of general circulation model output for Indian summer monsoon. *Meteorol. Appl.* **20**(3): 349–356, doi: 10.1002/met.1294.

Ajaaj AA, Mishra AK, Khan AA. 2016. Comparison of bias correction techniques for GPCC rainfall data in semi-arid climate. *Stoch. Environ. Res. Risk Assess.* **30**(6): 1659–1675, doi: 10.1007/s00477-015-1155-9.

Aryaputera AW, Yang D, Zhao L, Walsh WM. 2015. Very short-term irradiance forecasting at unobserved locations using spatio-temporal kriging. *Sol. Energ.* **122**: 1266–1278, doi: 10.1016/j.solener.2015.10.023.

Basist A, Bell GD, Meentemeyer V. 1994. Statistical relationships between topography and precipitation patterns. *J. Clim.* **7**(9): 1305–1315, doi: 10.1175/1520-0442(1994)007<1305:SRBTAP>2.0.CO;2.

Biau G, Zorita E, von Storch H, Wackernagel H. 1999. Estimation of precipitation by kriging in the EOF space of the sea level pressure field. *J. Clim.* **12**(4): 1070–1085, doi: 10.1175/1520-0442(1999)012<1070:EOPBKI>2.0.CO;2.

BMLFUW. 2016. Bundesministerium für Land und Forstwirtschaft, Umwelt und Wasserwirtschaft (BMLFUW), Abteilung IV/4 - Wasserhaushalt. http://ehyd.gv.at (accessed 29 February 2016).

Boer EP, de Beurs KM, Hartkamp AD. 2001. Kriging and thin plate splines for mapping climate variables. *Int. J. Appl. Earth Obs. Geoinf.* **3**(2): 146–154, doi: 10.1016/S0303-2434(01)85006-6.

Box GE, Cox DR. 1964. An analysis of transformations. *J. R. Stat. Soc. Ser. B Method.* **26**(2): 211–252.

Brier GW. 1950. Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **78**(1): 1–3, doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

CGIAR-CSI. 2016. SRTM 90m digital elevation database v4.1. http://srtm.csi.cgiar.org (accessed 29 February 2016).

Dabernig M, Mayr GJ, Messner JW, Zeileis A. 2016. Spatial ensemble post-processing with standardized anomalies. Working Papers, Atmospheric and Cryospheric Institute, University of Innsbruck. http://EconPapers.repec.org/RePEc:inn:wpaper:2016-08 (accessed 27 September 2016).

Daly C, Neilson RP, Phillips DL. 1994. A statistical-topographic model for mapping climatological precipitation over mountainous terrain. *J. Appl. Meteorol.* **33**(2): 140–158, doi: 10.1175/1520-0450(1994)033<0140:ASTMFM>2.0.CO;2.

Daly C, Taylor G, Gibson W. 1997. The PRISM approach to mapping precipitation and temperature. In *Proceeding: 10th AMS Conference on Applied Climatology*, Reno, NV, 208–209.

Daly C, Gibson WP, Taylor GH, Johnson GL, Pasteris P. 2002. A knowledge-based approach to the statistical mapping of climate. *Clim. Res.* **22**(2): 99–113, doi: 10.3354/cr022099.

Daly C, Halbleib M, Smith JI, Gibson WP, Doggett MK, Taylor GH, Curtis J, Pasteris PP. 2008. Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *Int. J. Climatol.* **28**(15): 2031–2064, doi: 10.1002/joc.1688.

Eden JM, Widmann M, Maraun D, Vrac M. 2014. Comparison of GCM- and RCM-simulated precipitation following stochastic post-processing. *J. Geophys. Res. Atmos.* **119**(19): 11040–11053, doi: 10.1002/2014JD021732.

Ekhart E. 1948. Die Niederschlagsverteilung in den Alpen nach dem Anomalienprinzip. *Geogr. Ann.* **30**: 728–739, doi: 10.2307/519914.

Fahrmeir L, Kneib T, Lang S, Marx B. 2013. *Regression – Models, Methods and Applications*. Springer-Verlag: Berlin. ISBN 978-3-642-34332-2.

Frei C, Schär C. 1998. A precipitation climatology of the Alps from high-resolution rain-gauge observations. *Int. J. Climatol.* **18**: 873–900, doi: 10.1002/(SICI)1097-0088(19980630)18:8<873::AID-JOC255>3.0.CO;2-9.

Gneiting T, Balabdaoui F, Raftery AE. 2007. Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc. Ser. B Stat. Method.* **69**(2): 243–268, doi: 10.1111/j.1467-9868.2007.00587.x.

Goovaerts P. 2000. Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. *J. Hydrol.* **228**(1–2): 113–129, doi: 10.1016/S0022-1694(00)00144-X.

Guan BT, Hsu H-W, Wey T-H, Tsao L-S. 2009. Modeling monthly mean temperatures for the mountain regions of Taiwan by generalized additive models. *Agr. Forest. Meteorol.* **149**(2): 281–290, doi: 10.1016/j.agrformet.2008.08.010.

Guisan A, Edwards TC Jr, Hastie T. 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecol. Model.* **157**(2–3): 89–100, doi: 10.1016/S0304-3800(02)00204-1.

Hong Y, Nix HA, Hutchinson MF, Booth TH. 2005. Spatial interpolation of monthly mean climate data for China. *Int. J. Climatol.* **25**(10): 1369–1379, doi: 10.1002/joc.1187.

Houze R. 2012. Orographic effects on precipitating clouds. *Rev. Geophys.* **50**(1): 1–47, doi: 10.1029/2011RG000365.

Hutchinson MF. 1998a. Interpolation of rainfall data with thin plate smoothing splines – Part II: analysis of topographic dependence. *J. Geogr. Inform. Decis. Anal.* **2**(2): 152–167.

Hutchinson MF. 1998b. Interpolation of rainfall data with thin plate smoothing splines – Part I: two dimensional smoothing of data

with short range correlation. *J. Geogr. Inform. Decis. Anal.* **2**: 168–185.
Hutchinson M. 2014. ANUSPLIN version 4.4. Centre for Resource and Environmental Studies. http://fennerschool.anu.edu.au/research/products/ (accessed 27 September 2016).
Jarvis CH, Stuart N. 2001. A comparison among strategies for interpolating maximum and minimum daily air temperatures. Part II: the interaction between number of guiding variables and the type of interpolation method. *J. Appl. Meteorol.* **40**(6): 1075–1084, doi: 10.1175/1520-0450(2001)040<1075:ACASFI>2.0.CO;2.
Klein N, Kneib T, Lang S, Sohn A. 2015. Bayesian structured additive distributional regression with an application to regional income inequality in Germany. *Ann. Appl. Stat.* **9**(2): 1024–1052, doi: 10.1214/15-AOAS823.
Messner JW, Mayr GJ, Wilks DS, Zeileis A. 2014. Extending extended logistic regression: extended versus separate versus ordered versus censored. *Mon. Weather Rev.* **142**(8): 3003–3014, doi: 10.1175/MWR-D-13-00355.1.
Price DT, McKenney DW, Nalder IA, Hutchinson MF, Kesteven JL. 2000. A comparison of two statistical methods for spatial interpolation of Canadian monthly mean climate data. *Agr. Forest. Meteorol.* **101**(2–3): 81–94, doi: 10.1016/S0168-1923(99)00169-0.
Rajczak J, Kotlarski S, Schär C. 2016. Does quantile mapping of simulated precipitation correct for biases in transition probabilities and spell lengths? *J. Clim.* **29**(5): 1605–1615, doi: 10.1175/JCLI-D-15-0162.1.
Raulin V. 1879. Über die Verteilung des Regens im Alpengebiet von Wien bis Marseille. *Zeitschrift der österreichischen Gesellschaft für Meteorologie* **14**: 233–247.
Rigby RA, Stasinopoulos DM. 2005. Generalized additive models for location, scale and shape, (with discussion). *Appl. Stat.* **54**: 507–554, doi: 10.1111/j.1467-9876.2005.00510.x.
Rust HW, Vrac M, Sultan B, Lengaigne M. 2013. Mapping weather-type influence on Senegal precipitation based on a spatial–temporal statistical model. *J. Clim.* **26**(20): 8189–8209, doi: 10.1175/JCLI-D-12-00302.1.
Sansom J, Tait A. 2004. Estimation of long-term climate information at locations with short-term data records. *J. Appl. Meteorol.* **43**(6): 915–923, doi: 10.1175/1520-0450(2004)043<0915:EOLCIA>2.0.CO;2.
Scheuerer M. 2014. Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Q. J. R. Meteorol. Soc.* **140**(680): 1086–1096, doi: 10.1002/qj.2183.
Scheuerer M, Büermann L. 2014. Spatially adaptive post-processing of ensemble forecasts for temperature. *J. R. Stat. Soc. Ser. C Appl. Stat.* **63**(3): 405–422, doi: 10.1111/rssc.12040.
Scheuerer M, Hamill TM. 2015. Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Mon. Weather Rev.* **143**(11): 4578–4596, doi: 10.1175/MWR-D-15-0061.1.
Snepvangers J, Heuvelink G, Huisman J. 2003. Soil water content interpolation using spatio-temporal kriging with external drift. *Geoderma* **112**(3–4): 253–271. Pedometrics 2001 doi: 10.1016/S0016-7061(02)00310-5.
Stauffer R, Messner JW, Mayr GJ, Umlauf N, Zeileis A. 2016. Ensemble post-processing of daily precipitation sums over complex terrain using censored high-resolution standardized anomalies. Working papers, Faculty of Economics and Statistics, University of Innsbruck. http://EconPapers.repec.org/RePEc:inn:wpaper:2016-21 (accessed 27 September 2016).
Stidd CK. 1973. Estimating the precipitation climate. *Water Resour. Res.* **9**(5): 1235–1241, doi: 10.1029/WR009i005p01235.
Themeßl MJ, Gobiet A, Heinrich G. 2012. Empirical-statistical downscaling and error correction of regional climate models and its impact on the climate change signal. *Clim. Change* **112**(2): 449–468, doi: 10.1007/s10584-011-0224-4.
Thiessen AH. 1911. Precipitation averages for large areas. *Mon. Weather Rev.* **39**(7): 1082–1089, doi: 10.1175/1520-0493(1911)39<1082b:PAFLA>2.0.CO;2.
Thornton PE, Running SW, White MA. 1997. Generating surfaces of daily meteorological variables over large regions of complex terrain. *J. Hydrol.* **190**(3–4): 214–251, doi: 10.1016/S0022-1694(96)03128-9.
Umlauf N, Klein N, Zeileis A. 2016a. *bamlss: Bayesian additive models for location, scale and shape (and beyond). Unpublished manuscript*.
Umlauf N, Zeileis A, Klein N, Adler D. 2016b. *bamlss: Bayesian additive models for location scale and shape (and beyond)*. https://R-Forge.R-project.org/R/?group_id=865 (accessed 27 September 2016).
Vicente-Serrano SM, Angel Saz-Sánchez M, Cuadrat JM. 2003. Comparative analysis of interpolation methods in the middle Ebro Valley (Spain): application to annual precipitation and temperature. *Clim. Res.* **24**(2): 161–180, doi: 10.3354/cr024161.
Wapler K. 2013. High-resolution climatology of lightning characteristics within Central Europe. *Meteorol. Atmos. Phys.* **122**(3–4): 175–184, doi: 10.1007/s00703-013-0285-1.
Wong G, Maraun D, Vrac M, Widmann M, Eden JM, Kent T. 2014. Stochastic model output statistics for bias correcting and downscaling precipitation including extremes. *J. Clim.* **27**(18): 6940–6959, doi: 10.1175/JCLI-D-13-00604.1.
Wood SN. 2003. Thin plate regression splines. *J. R. Stat. Soc. Ser. B Stat. Method.* **65**(1): 95–114, doi: 10.1111/1467-9868.00374.
Wood S. 2006. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC: Boca Raton (FL), London, New York. ISBN: 978-1498728331.
Yee T. 2015. *Vector Generalized Linear and Additive Models: With an Implementation in R*. Springer: New York, NY.

Article III

Stauffer R., Umlauf N., Messner J.W., Mayr G.J., and Zeileis A. (2017). *Ensemble Postprocessing of Daily Precipitation Sums over Complex Terrain Using Censored High-Resolution Standardized Anomalies*. Monthly Weather Review, 145, 955–969, doi:10.1175/MWR-D-16-0260.1.

JCR ranking: **Category 1** in *Meteorology & Atmospheric Sciences*.

Ensemble Postprocessing of Daily Precipitation Sums over Complex Terrain Using Censored High-Resolution Standardized Anomalies

RETO STAUFFER

Department of Statistics, and Institute of Atmospheric and Cryospheric Sciences, University of Innsbruck, Innsbruck, Austria

NIKOLAUS UMLAUF

Department of Statistics, University of Innsbruck, Innsbruck, Austria

JAKOB W. MESSNER

Department of Statistics, and Institute of Atmospheric and Cryospheric Sciences, University of Innsbruck, Innsbruck, Austria

GEORG J. MAYR

Institute of Atmospheric and Cryospheric Sciences, University of Innsbruck, Innsbruck, Austria

ACHIM ZEILEIS

Department of Statistics, University of Innsbruck, Innsbruck, Austria

(Manuscript received 16 July 2016, in final form 18 October 2016)

ABSTRACT

Probabilistic forecasts provided by numerical ensemble prediction systems have systematic errors and are typically underdispersive. This is especially true over complex topography with extensive terrain-induced small-scale effects, which cannot be resolved by the ensemble system. To alleviate these errors, statistical post-processing methods are often applied to calibrate the forecasts. This article presents a new full-distributional spatial postprocessing method for daily precipitation sums based on the standardized anomaly model output statistics (SAMOS) approach. Observations and forecasts are transformed into standardized anomalies by subtracting the long-term climatological mean and dividing by the climatological standard deviation. This removes all site-specific characteristics from the data and makes it possible to fit one single regression model for all stations at once. As the model does not depend on the station locations, it directly allows the creation of probabilistic forecasts for any arbitrary location. SAMOS uses a left-censored power-transformed logistic response distribution to account for the large fraction of zero observations (dry days), the limitation to nonnegative values, and the positive skewness of the data. ECMWF reforecasts are used for model training and to correct the ECMWF ensemble forecasts with the big advantage that SAMOS does not require an extensive archive of past ensemble forecasts as only the most recent four reforecasts are needed, and it automatically adapts to changes in the ECMWF ensemble model. The application of the new method to the central Alps shows that the new method is able to depict the small-scale properties and returns accurate fully probabilistic spatial forecasts.

Denotes content that is immediately available upon publication as open access.

Corresponding author e-mail: Reto Stauffer, reto.stauffer@uibk.ac.at

DOI: 10.1175/MWR-D-16-0260.1

1. Introduction

In mountainous regions, large amounts of precipitation can lead to severe floods and landslides during spring and summer and to dangerous avalanche conditions during winter. Accurate and reliable knowledge about the expected precipitation can therefore be crucial for strategic planning and to raise awareness among the public.

Precipitation forecasts, or weather forecasts in general, are typically provided by numerical weather prediction models. Nowadays, most forecast centers also compute probabilistic forecasts based on numerical ensemble prediction systems (EPSs; Epstein 1969; Buizza et al. 2005) as probabilistic information can be crucial, for example, for strategic planning or decision-makers. An ensemble consists of several (independent) forecast runs with slightly different initial conditions, model physics, and/or parameterizations. The goal of an EPS is not only to provide one single forecast but also to provide additional information about the weather-situation-dependent forecast uncertainty. Although EPSs are undergoing constant improvements, they are not able to provide fully reliable forecasts and are typically underdispersive (Mullen and Buizza 2001; Hagedorn et al. 2012).

To correct for systematic errors and to correct the uncertainty provided by the EPS, postprocessing methods are often applied. A variety of ensemble postprocessing methods for precipitation are available nowadays, such as analog methods (Hamill et al. 2006, 2015), ensemble dressing (Roulston and Smith 2003), Bayesian model averaging (BMA; Sloughter et al. 2007; Fraley et al. 2010), extended logistic regression (Wilks 2009; Ben Bouallègue and Theis 2014; Messner et al. 2014b), or nonhomogeneous regression (Gneiting et al. 2005). Several extensions exist for nonnormally distributed variables (Thorarinsdottir and Gneiting 2010; Lerch and Thorarinsdottir 2013; Scheuerer 2014; Scheuerer and Hamill 2015). For precipitation, Messner et al. (2014a) show that a censored logistic regression fits well, while Scheuerer (2014) and Scheuerer and Hamill (2015) use a left-censored generalized extreme value (GEV) distribution or a left-censored shifted gamma distribution, respectively.

These postprocessing methods are often applied on a station or gridpoint level such that for each location, one set of regression coefficients is estimated to correct the ensemble forecasts. However, for a wide range of applications, predictions for locations between observational sites are of great interest. Therefore, the regression models have to be extended such that spatial probabilistic predictions can be made.

In this article, a new spatial statistical postprocessing method for daily precipitation sums over complex terrain is presented. Even on a small spatial scale, two neighboring stations can show very different characteristics in terms of observed precipitation sums. These differences can be caused by topographically induced flow regimes, orographic lifting and shading effects, convective regimes, and many other factors. Most of these processes cannot yet be resolved by global EPS models. To account for these small-scale spatial variabilities among all stations, we are using an adapted version of the anomaly approach first published by Scheuerer and Büermann (2014) and further extended by Dabernig et al. (2017). Observations and ensemble forecasts are transformed into standardized anomalies by subtracting the long-term climatological mean and dividing by the climatological standard deviation. This removes the station-dependent characteristics from the data and makes it possible to fit one single regression model for all stations at once. As the model does not rely on site-specific characteristics anymore, the corrections can be applied to future ensemble forecasts to create probabilistic forecasts for any arbitrary location within the area of interest.

Following Dabernig et al. (2017), we use the standardized anomaly model output statistics (SAMOS) approach and extend the framework to fulfill all requirements needed for precipitation postprocessing. SAMOS offers a simple and computationally efficient framework for fully probabilistic spatial postprocessing and is applied to the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble in combination with the ECMWF reforecasts. The approach presented qualifies for an operational system as no extensive archive of historical forecasts is required. SAMOS uses a rolling 4-week time window as a training dataset so that only the reforecasts of the most recent month from the operational ECMWF data dissemination have to be retained, which currently (in 2016) consist of eight independent reforecast runs covering the previous 20 years. Because of this rolling training dataset, SAMOS automatically adapts itself to the latest ensemble model version within a very short time period.

2. Area of interest and data

a. Study area

To develop and validate the new method presented in this study, we focus on the governmental area of Tyrol, Austria. Tyrol has a size of about 12 500 km^2 and is home to approximately 740 000 inhabitants (Statistik Austria 2016) living in the two separated parts, with North Tyrol on the north side of the main Alpine ridge and East Tyrol south of the main Alpine ridge. The study area is located in the eastern part of the Alps, showing a highly complex topography. Figure 1 shows the state borders of Tyrol and the topography reaching from 465 to 3798 m MSL, including some of the highest mountains in

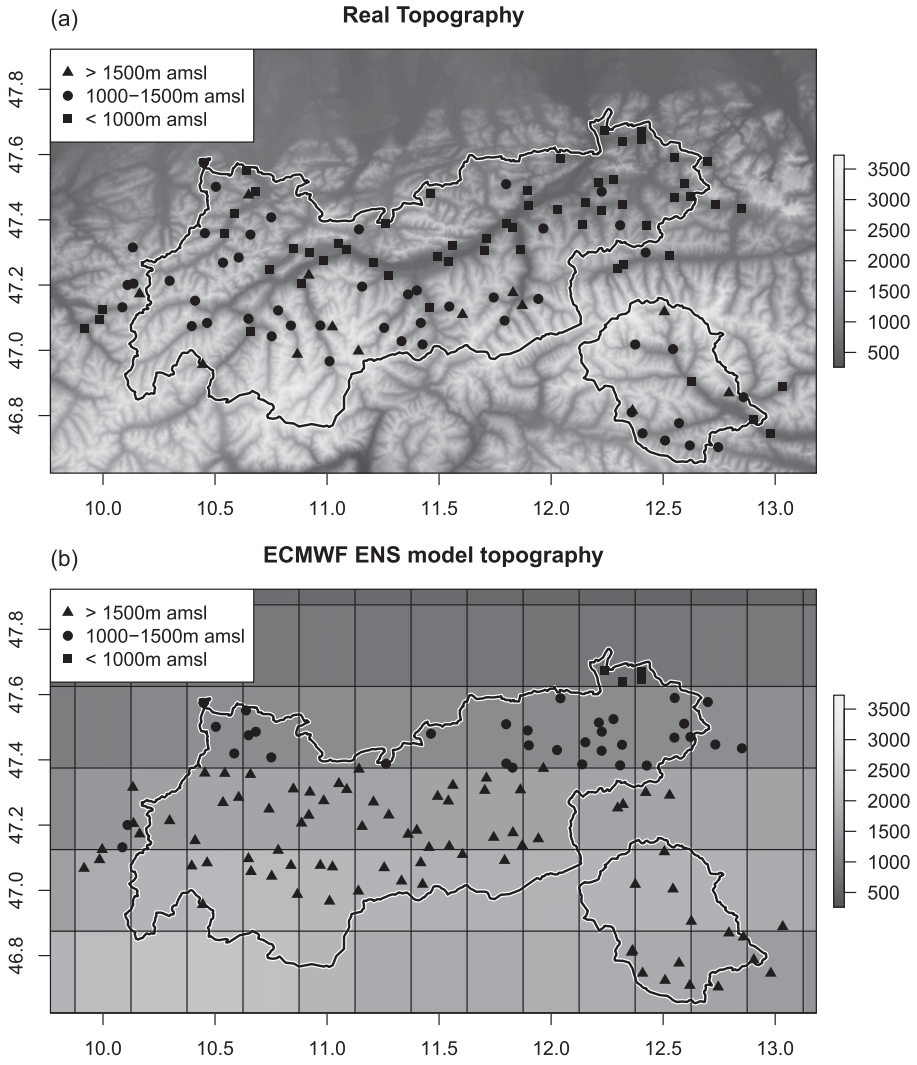

FIG. 1. The black line shows the state borders of Tyrol. Each marker represents an observation site (total 117), the marker type indicates the altitude: square (≤1000 m MSL), bullet (1000–1500 m MSL), and triangle (≥1500 m MSL) with respect to the underlying topography. The background shows the (a) real topography (Jarvis et al. 2008) and the (b) ECMWF EPS model topography with a 0.5° resolution as used between February 2010 and December 2012.

Austria. Because of the high population density and the strong economic focus on tourism (>10 million tourists in 2014; Kaiser et al. 2014), there is a high demand for accurate weather forecasts.

b. *Observational data*

The local hydrographical service provides a dense precipitation measurement network, whereof 117 stations in Tyrol and its surroundings will be used for model training and validation spanning September 1971 through the end of 2012. The mean distance to the four closest stations in the surroundings is only about 10 km. Locations of the observation sites are highlighted in Fig. 1. The hydrographical service performs rigorous quality controls on the observations and makes them freely available for any noncommercial use on the maintainers' website (Bundesministerium für Land und Forstwirtschaft, Umwelt und Wasserwirtschaft 2016).

c. *Numerical weather forecast data*

The numerical forecasts are obtained from the ECMWF, including the operational ensemble (ENS; 0000 UTC initial), which consists of 50 + 1 individual forecasts based on perturbed initial conditions (50 forecasts plus control run) and the ECMWF reforecast dataset. The ECMWF reforecast dataset has existed since 18 February 2010 and was slightly extended over the years. Until 14 June 2012, the reforecast was computed once a week, providing ensemble reforecasts consisting of 4 + 1 members for the most recent 18 years.

From 21 June 2012 through the end of 2012, the number of years was extended from 18 to 20. This reforecast is designed to provide the model climate of the latest ECMWF ENS version and is often used for model calibration (e.g., Hamill et al. 2008; Hamill 2012).

In this study, the time period from February 2010 to December 2012 is used. Every Thursday, the reforecasts for the same date two weeks in advance have been computed, including a 4 + 1 member ensemble for the most recent 18–20 years. As an example, on Thursday 1 November 2012, the reforecast for 15 November has become available for the most recent 20 years, namely 15 November 2011, 15 November 2010, . . . , 15 November 1992, with 4 + 1 members each.

d. Training and verification dataset

The ECMWF reforecasts are used to compute the climatology of the ECMWF ensemble, which will be used as background information and to train the statistical postprocessing, including the most recent four reforecast runs centered around the current date (computed every Thursday; section 2c). Therefore, the model climatology is based on 4 runs $\times$ 5 members $\times$ 20 yr = 400 individual forecasts (details in section 3c). For the training dataset, the reforecasts are bilinearly interpolated to each of the 117 observation sites. Out of each interpolated reforecast ensemble (daywise, 4 + 1 members), the mean and standard deviation is used later to build the training dataset. We use the most recent four 0000 UTC reforecast runs yielding to a training sample of up to 4 runs $\times$ 20 yr = 80 data pairs per station, or 4 runs $\times$ 20 yr $\times$ 117 stations = 9360 observation–reforecast pairs for the full spatial SAMOS (details in section 4b).

Once the regression coefficients are estimated, the correction can be applied to future EPS forecasts using the mean and standard deviation of the 50 + 1 members of the ECMWF ENS.

Because of the availability of the observations (section 2b) and the ECMWF reforecasts (section 2c), the time period between 26 February 2010 and 31 December 2012 will be used for verification, with an overall data availability of 99.4% and roughly 120 500 unique observation–forecast pairs.

3. Methodology

a. Censored nonhomogeneous logistic regression (CNLR)

The distribution of precipitation observations at a particular observation site shows three main properties: it is limited to nonnegative values, has a large fraction of 0 observations (dry days), and is strongly positively skewed. We take the nonhomogeneous Gaussian regression (NGR; Gneiting et al. 2005) as our base model and extend the NGR framework to suit spatial precipitation postprocessing.

In contrast to the original NGR, a *logistic* response distribution is assumed. The logistic distribution shows a similar bell shape as the Gaussian distribution but has slightly heavier tails. The logistic distribution is defined by two parameters: the *location* μ describing the mean and the *scale* σ describing the width of the distribution. To remove the positive skewness, a power transformation $1/p$ is applied to the observations and to every ensemble member (Box and Cox 1964). Different power parameters p have already been suggested in the literature for precipitation applications such as $p = 3$ (Stidd 1973) or $p = 2$ (Hutchinson 1998). However, the optimal power parameter is a function of the data, the model assumptions, and the application. For this study, the power parameter p has been set to $p = 1.35$, which turned out to fit best for the dataset and distribution used (details in section 3c).

Furthermore, the response is assumed to be left censored at 0 to account for the nonnegative observations and the large fraction of 0 observations. The concept of left censoring assumes that there is an underlying *latent* (unobservable) process driving the observable response, which can be described by a linear predictor. While the latent response y is allowed to become negative, the observable response "precipitation" is simply 0 if the latent response y is below zero or the inverse power-transformed latent response y^p otherwise. For simplicity, the zero left-censored nonhomogeneous logistic regression will be denoted as CNLR from now on.

Both distributional parameters (μ, σ) are expressed by a linear predictor including the covariates or explanatory variables. As suggested by Gneiting et al. (2005), the mean of the ensemble forecast drives the location μ, and the standard deviation of the ensemble drives the scale σ. For this study, we only use the forecasted daily accumulated total precipitation from the ensemble (section 2c) as the meteorological predictor variable. In Eq. (1), m denotes the mean, and s denotes the standard deviation of the forecasted power-transformed daily total precipitation amounts of the ensemble members.

Following the idea of Gebetsberger et al. (2016), a second covariate z has been included. The term z is a binary split variable, which takes 1 if all forecast members in the training dataset predict less than 0.01 mm day^{-1} ($z = 1$; "no" precipitation) or 0 otherwise. This allows us to handle dry and wet cases differently and has a positive impact on the results. It furthermore solves the problem of taking the logarithm of the ensemble standard deviation if all members predict 0 mm, which leads to $s = 0$. The log transformation

on the scale σ is used to ensure nonnegative-scale values during optimization. The full CNLR assumptions can then be written as

$$\begin{aligned}\text{precipitation} &= \begin{cases} 0 & \text{if } y \le 0 \\ y^p & \text{else} \end{cases}, \\ y &\sim \mathscr{L}(\mu, \sigma), \\ \mu &= \beta_0 + \beta_1 \times z + \beta_2 \times m \times (1 - z), \\ \log(\sigma) &= \gamma_0 + \gamma_1 \times \log(s) \times (1 - z). \end{aligned} \tag{1}$$

In case of a dry ensemble forecast ($z = 1$), the linear predictors collapse to $\mu = \beta_0 + \beta_1$ and $\log(\sigma) = \gamma_0$ such that the model only consists of two estimated constants describing the climatological distribution of the response conditional on all cases where $z = 1$. The variable β_1 typically becomes strongly negative, which leads to a strongly negative latent location μ and overall small expected amounts of precipitation for the case $z = 1$. For wet cases ($z = 0$), the linear predictors become $\mu = \beta_1 + \beta_2 \times m$ and $\log(\sigma) = \gamma_0 + \gamma_1 \times \log(s)$, which corresponds to the NGR model proposed by Gneiting et al. (2005). These assumptions allow us to correct the bias but also a possible overdispersion or underdispersion of the ensemble as the scale σ depends on the predicted ensemble standard deviation. Even if the two cases are not independent and connected via the scale part, discontinuities occur at the transition where z goes from 0 to 1. As this only happens in regions with very small predicted amounts of precipitation, the effect on the results is marginal.

The model as specified in Eq. (1) can be applied at every arbitrary location where both historical observations and historical ensemble forecasts are available. For pointwise ensemble postprocessing, *one CNLR model* has to be fitted at *each observation site*. In this case, all CNLR models are independent and have their own regression coefficients $\beta_\bullet$ and $\gamma_\bullet$. As these coefficients are site specific, spatial predictions are not directly possible and would require an additional interpolation method, which allows us to account for supplementary covariates, such as terrain or surface properties.

Instead of a two-step approach of performing stationwise estimates and interpolating/extrapolating the resulting coefficients afterward, we extend the model to include the training data of all stations at once and fit one simple and computationally efficient model for fully probabilistic spatial estimates.

b. SAMOS

The statistical method presented in this article is based on the anomaly approach first published by Scheuerer and Büermann (2014) and further extended by Dabernig et al. (2017), focusing on temperature forecasts across Germany and northern Italy, respectively. We extend the SAMOS approach by Dabernig et al. (2017), yielding to a censored SAMOS version for precipitation postprocessing.

Climatological properties between two precipitation observation sites may vary in mean (location) and variability (scale). This is especially true over complex terrain where only a few kilometers between a valley and a mountain station can result in very large climatological differences (Frei and Schär 1998; Isotta et al. 2014; Stauffer et al. 2017). These small-scale features influence daily precipitation sums but are not yet fully resolved by global numerical ENSs. Therefore, a high-resolution *spatiotemporal climatology* is used as *background information* to provide small-scale features at any location within the study area. Instead of modeling the relationship between past observations and past numerical weather forecasts directly, the statistical model uses high-resolution *standardized anomalies*. Anomalies are defined as the short-term deviation from the local long-term climate. These anomalies can be divided by the local climatological variability to obtain standardized anomalies. Standardized anomalies of the observations (precipitation) are defined as

$$y^* = \frac{\text{precipitation}^{1/p} - \mu_{\text{obs,clim}}}{\sigma_{\text{obs,clim}}}, \tag{2}$$

where $\mu_{\text{obs,clim}}$ and $\sigma_{\text{obs,clim}}$ describe the long-term climatological properties of daily observations and will be discussed in detail in section 3c. The term y^* denotes the resulting latent response on the standardized anomaly scale, which follows a standard logistic distribution $\mathscr{L}(0, 1)$. Equivalent to Eq. (2), standardized anomalies of the ensemble forecasts (ens) can be computed with the climatological properties $\mu_{\text{ens,clim}}$ and $\sigma_{\text{ens,clim}}$ of the ensemble using

$$\text{ens}^* = \frac{\text{ens}^{1/p} - \mu_{\text{ens,clim}}}{\sigma_{\text{ens,clim}}}. \tag{3}$$

The ensemble climatology ($\mu_{\text{ens,clim}}$, $\sigma_{\text{ens,clim}}$) is described in section 3c.

Because of standardization, the censoring point on the anomaly scale becomes a function of the observed climatology. While the censoring point is on 0 (no precipitation) on the original or power-transformed scale [Eq. (1)], the censoring threshold becomes $-\mu_{\text{obs,clim}} / \sigma_{\text{obs,clim}}$ after standardizing the data. Figure 2a shows the power-transformed observations with a constant censoring threshold of 0 throughout the whole year. Figure 2b shows all standardized anomalies and

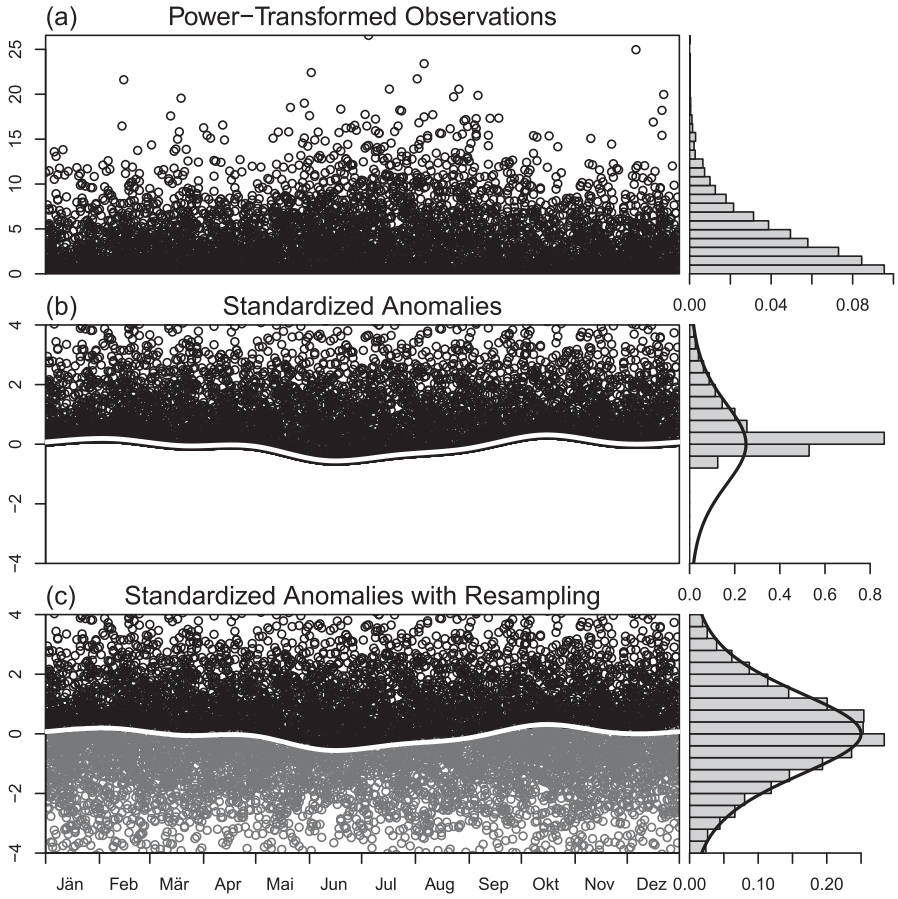

FIG. 2. Example of standardized anomalies for one specific station (Bromberg, Austria) with roughly 8500 unique daily observations between 1987 and 2013. (a) Daily observations on power-transformed scale ($\mathrm{mm}^{1/p}\,\mathrm{day}^{-1}$ with $p = 1.35$); (b) standardized anomalies; and (c) standardized anomalies with simulated censored data (for visual justification only). (left) Data plotted against the day of the year. The solid white lines in (b) and (c) show the shifted censoring point due to standardization. Simulated censored observations are shown in gray in (c). (right) Density histograms, with the standard logistic distribution shown as solid lines [$\mathscr{L}(0, 1)$; (b) and (c)].

the shifted censoring threshold indicated by the solid line. As observations below the censoring threshold never occur, all data points lie on or above this line. Figure 2c is an extension of Fig. 2b, where all observations on the censoring threshold ($0\,\mathrm{mm\,day}^{-1}$ on the original scale) were simulated from the standard logistic distribution for visual justification. As shown in the density plot, the standardized anomalies now follow a latent standard logistic distribution $\mathscr{L}(0, 1)$. As each of the 117 stations is standardized using its specific climatological properties $\mu_{\mathrm{obs,clim}}$ and $\sigma_{\mathrm{obs,clim}}$, the standardized anomalies of all stations show the same distribution [$\mathscr{L}(0, 1)$]. Thus, the standardization removes site-specific features from the data and brings the data of all stations onto a comparable level.

Combining the CNLR model from Eq. (1) with the concept of standardized anomalies [Eqs. (2) and (3)] leads to the full specification of the SAMOS model with a left-censored logistic response:

$$\begin{aligned} y^* &\sim \mathscr{L}(\mu^*, \sigma^*), \\ \mu^* &= \beta_0 + \beta_1 \times z + \beta_2 \times m^* \times (1 - z), \\ \log(\sigma^*) &= \gamma_0 + \gamma_1 \times \log(s^*) \times (1 - z). \end{aligned} \tag{4}$$

As on the power-transformed scale, the standardized anomalies y^* are still assumed to follow a logistic distribution. The linear predictors for location μ^* and $\log(\sigma^*)$ on the standardized anomaly scale depend on the standardized ensemble anomalies (ens*) and the binary split indicator z. In this study, total precipitation forecasts are used as the only meteorological variable. The covariates m^* and s^* therefore correspond to the empirical mean and standard deviation of the standardized total precipitation forecast anomalies [Eq. (3)].

Once all covariates are known, the regression coefficients of the SAMOS model given by Eqs. (2)–(4) can be estimated using censored maximum likelihood optimization as offered by the R package crch (Messner et al. 2016) or similar software. The climatological estimates required to create the standardized anomalies are explained in detail in section 3c.

Given all regression parameters $\beta_\bullet$ and $\gamma_\bullet$ of the SAMOS model [Eq. (4)], the correction can be applied to future ensemble forecasts. As the SAMOS model returns both parameters on the standardized anomaly scale, they have to be destandardized with respect to the spatial climatology:

$$\mathscr{L}_0\left(\underbrace{\mu^* \times \sigma_{\text{obs,clim}} + \mu_{\text{obs,clim}}}_{\text{location}}, \underbrace{\sigma^* \times \sigma_{\text{obs,clim}}}_{\text{scale}}\right). \tag{5}$$

The destandardized zero left-censored distribution $\mathscr{L}_0(\cdots)$ describes the full postprocessed ENS forecast distribution on the power-transformed scale. Since the SAMOS regression coefficients are location independent, the postprocessed predictions can be computed at any location within the study area where both ENS forecasts and climatological estimates ($\mu_{\bullet,\text{clim}}, \sigma_{\bullet,\text{clim}}$) are available. As spatiotemporal climatologies are used (details in section 3c), the only limitation for the postprocessed ENS forecasts is the horizontal grid spacing of the spatial climatology, which itself only depends on the resolution of the available digital elevation model (see Stauffer et al. 2017). From the full-probabilistic SAMOS forecasts, different properties can then be derived, such as the mean or expectation, quantiles, probability of precipitation, or probabilities exceeding a certain threshold. To retrieve the corrected forecasts on the original scale in millimeters per day, the inverse power transformation has to be taken into account. Details can be found in appendix A.

In the limiting case that the ensemble would not provide any information at all, μ^* approaches 0 and σ^* approaches 1, resulting in $\mu = \mu_{\text{obs,clim}}$ and $\sigma = \sigma_{\text{obs,clim}}$, which corresponds to the underlying high-resolution climatology—the most reliable information available in this case.

c. *Climatological estimates*

The climatological properties $\mu_{\bullet,\text{clim}}$ and $\sigma_{\bullet,\text{clim}}$ for both the observations and the ensemble forecasts have to be specified to be able to derive the standardized anomalies y^* and ens* [Eqs. (2) and (3)]. The computation of the observed climatology is based on Stauffer et al. (2017) but uses a left-censored logistic instead of Gaussian distribution and consequently a modified power-transformation parameter. As in Stauffer et al. (2017), the optimal power parameter was chosen using a power-adjusted maximum likelihood approach optimizing 117 stationwise climatologies. Since the optimal power parameters did not show a distinct spatial or altitudinal dependency, the median among all 117 estimates was selected using a constant $p = 1.35$ in this study.

The observed spatiotemporal climatology is based on all 117 stations (Fig. 1) and uses daily precipitation measurements from 1971 through the end of 2009, yielding to roughly 1.5 million individual observations. Data from the years 2010–13 are set aside for verification.

The climatology is based on a nonhomogeneous regression model similar to the SAMOS method. In contrast to Eqs. (1) and (4), the linear predictors of the climatological model include smooth one-dimensional and multidimensional spline effects to depict all features of the climatology. In addition to the global intercepts (β, γ), an altitudinal effect (s_1, t_1), an effect to describe the seasonality based on the day of the year (s_2, t_2), a spatial effect on dependent longitude and latitude (s_3, t_3), and a three-dimensional effect to describe spatial variations in the seasonal pattern (s_4, t_4) are included. Further details can be found in Stauffer et al. (2017). The full model specification of the observation climatology can be expressed as

$$\begin{aligned}
\text{precipitation} &= \begin{cases} 0 & \text{if} \quad y \le 0 \\ y^p & \text{else} \end{cases}, \\
y &\sim \mathscr{L}(\mu_{\text{obs,clim}}, \sigma_{\text{obs,clim}}), \\
\mu_{\text{obs,clim}} &= \beta + s_1(\text{alt}) + s_2(\text{yday}) \\
&\quad + s_3(\text{lon, lat}) + s_4(\text{yday, lon, lat}), \\
\log(\sigma_{\text{obs,clim}}) &= \gamma + t_1(\text{alt}) + t_2(\text{yday}) \\
&\quad + t_3(\text{lon, lat}) + t_4(\text{yday, lon, lat}).
\end{aligned} \tag{6}$$

Again, both parameters of the power-transformed left-censored logistic distribution (location $\mu_{\text{obs,clim}}$ and scale $\sigma_{\text{obs,clim}}$) are modeled. This is required as they are used for the standardization of the SAMOS model. Although the climatology model (section 3c) is quite complex, estimation only takes about 30 h and has to be done rarely, for example, once a year.

In addition to climatological estimates of the observations, climatological estimates $\mu_{\text{ens,clim}}$ and $\sigma_{\text{ens,clim}}$ are required to compute standardized anomalies of the ensemble forecasts as in Eq. (3). The two parameters represent the long-term climatology of the ECMWF EPS (section 2c) and are computed from the ECMWF reforecast dataset. The mean and standard deviation are

based on up to 400 individual forecasts provided by the most recent four reforecast runs (section 2d):

$$\mu_{\text{ens,clim}} = \text{mean(reforecast)},$$
$$\sigma_{\text{ens,clim}} = \text{stdv(reforecast)} \times C. \quad (7)$$

The climatological location $\mu_{\text{ens,clim}}$ is simply the empirical mean; the climatological scale $\sigma_{\text{ens,clim}}$ is the ''standard deviation'' of the reforecast used. The factor $C = \sqrt{3}/\pi$ is used to get the empirical scale of a logistic distribution to be on the same scale as the estimated scale of the observation climatology $\sigma_{\text{obs,clim}}$ [Eq. (6)].

4. Results and verification

a. SAMOS results

Figure 3 shows an example of the climatologies used for 18 May 2010 and the resulting spatial SAMOS predictions. It can be seen in all climatological estimates (Figs. 1a–d) that the altitudinal dependency is the most dominant effect for this day (cf. Fig. 1). The ENS with a horizontal grid spacing of ~40 km × 40 km is only able to resolve the main Alpine ridge leading to the smooth north–south transitions in the left column of Fig. 3. The ensemble climatology correctly shows larger location μ (Fig. 3a) and scale σ (Fig. 3c) toward the prealpine flatland to the north and the south; however, this is only a very rough approximation of what is actually observed (Figs. 3b,d).

Figures 3e–h show the predictions for 18 May 2010, when a cold front hit the Alps from the north driven by a strongly pronounced low pressure system east of the study area. As a result, the forecasts show larger precipitation amounts north of the area due to orographic lifting and blocking. As the ENS is only able to represent the topography as one smooth ridge (Fig. 1), the only feature that can be identified in the ENS prediction is a gradual decrease of precipitation from north to south over the main Alpine ridge. In reality, a first mountain ridge alongside the northern boundary of the study area is blocking the air mass. Larger amounts of precipitation are typically observed in southern Germany north of Tyrol, while the well-marked Alpine valleys in Tyrol typically receive less precipitation. This can be seen in the observed climatology (Fig. 3b) but also for this particular day in the corrected SAMOS forecasts (Figs. 3f,h). South of the largest valley with a west–east orientation, increased forecasted amounts and probabilities can be seen in the corrected SAMOS predictions related to a secondary lifting of the air masses at the high mountains close to the main Alpine ridge.

The example shows that SAMOS is able to add interpretable and meaningful features to the ENS during the postprocessing procedure. However, the performance cannot be evaluated with a single case alone. Section 4b therefore contains a detailed analysis and verification on a 3-yr independent dataset.

b. Verification

For verification, the predictions of four different methods will be compared with unused (out of sample) data between February 2010 and December 2012. As two baseline methods, the climatologies (CLIM; section 3c) and the raw total precipitation predictions from the ECMWF ENS will be used. The empirical frequency of the 50 + 1 ensemble members is used as probability to compute the Brier scores shown in the results. Furthermore, a stationwise postprocessing (STN) is included based on Eq. (1). For STN, a separate CNLR model is estimated for each of the 117 stations in the dataset.

The predictions of all methods are out of sample such that the data used for verification are not included in the training dataset, which is used to estimate the regression coefficients. CLIM is based on all available observations except that the years 2010–13 are excluded (section 3c). Therefore, CLIM predictions are spatially in sample but temporally out of sample. STN is using the latest four available reforecast runs yielding to spatially in-sample but temporally out-of-sample predictions. SAMOS is the only method whose predictions can be verified both spatially and temporally out of sample. Therefore, a leave-one-out cross validation is performed. For each station, the SAMOS regression coefficients were estimated based on the most recent four reforecast runs excluding this one specific station. Forecasts were then made for the excluded station only. Table 1 contains a summary of all four methods and shows their sample behavior. Full in-sample SAMOS results are omitted as hardly any differences can be seen compared to the cross-validated out-of-sample results.

The continuous ranked probability score (CRPS; appendix B) of all predictions is shown as a continuous ranked probability skill score (CRPSS) in Fig. 4 using CLIM as reference. Values below zero indicate less predictive skill than CLIM. The higher the score, the better the performance of the corresponding method. As the CRPS is a fully probabilistic score, it penalizes for a possible dislocation of the predicted distribution but also for the wrongly predicted width or sharpness. The scores show an overall decrease with increasing forecast horizon for all three methods, slowly approaching the skill of the climatology. The two postprocessing methods STN and SAMOS show a significant

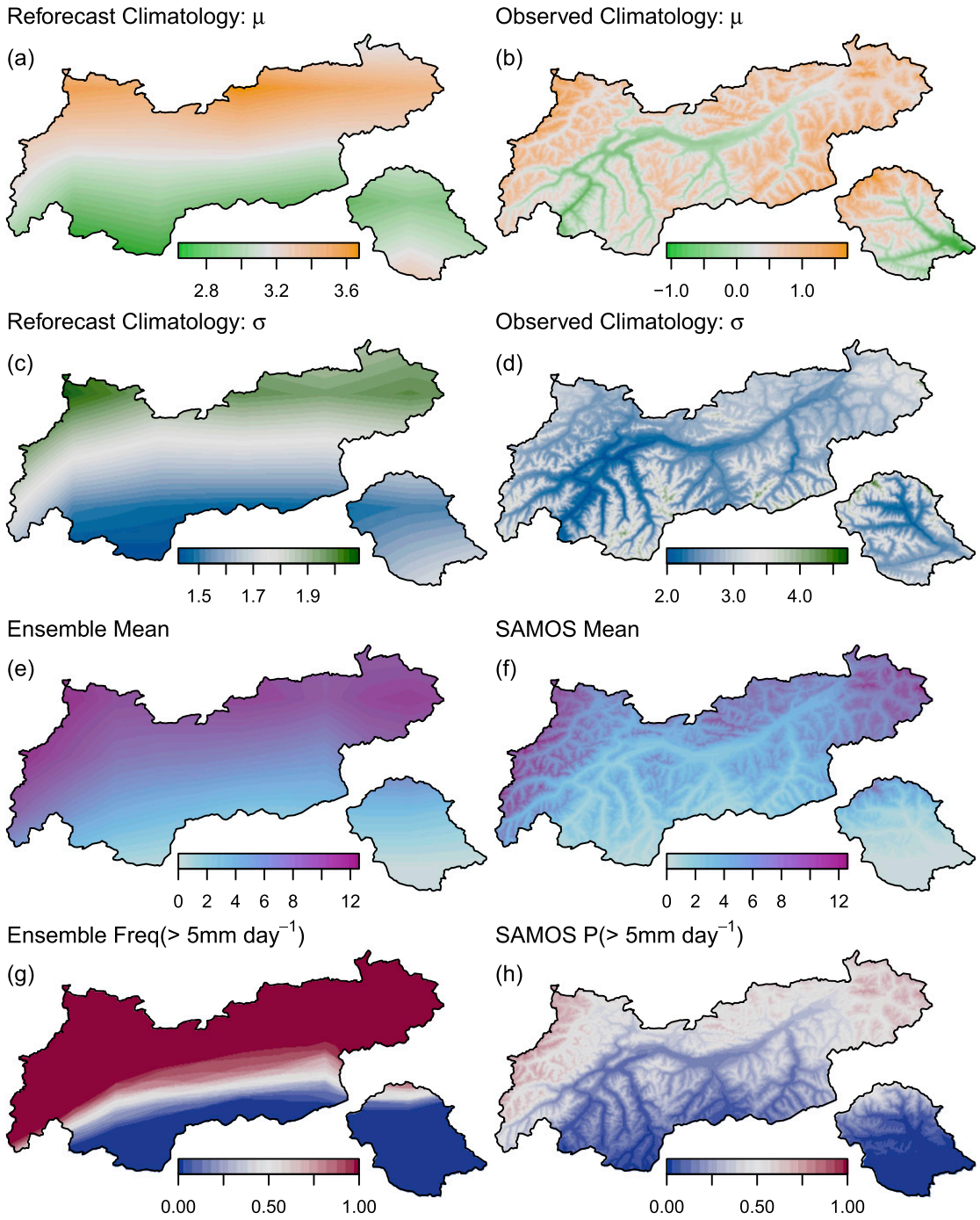

FIG. 3. Example prediction for 18 May 2010, 1-day-ahead forecast. (a),(b) Climatological location μ; (c),(d) climatological scale σ; (e),(f) forecast mean; and (g) frequency and (h) probability of exceeding 5 mm day^{-1}. (left) Reforecast climatologies and the raw ensemble forecast; (right) observed climatology and the postprocessed SAMOS predictions. Location μ and scale σ on the latent power-transformed scale are in mm$^{1/p}$ with $p = 1.35$. Note that (a),(b) and (c),(d) use different color scales regarding the range of the data.

improvement with respect to the ENS up to the 6-day-ahead forecasts. SAMOS outperforms the STN method, even if it is verified fully out of sample. The differences between STN and SAMOS are small but all significant (paired two-sided t test, 5% significance level; not shown).

In addition to the CRPSS, Fig. 5 shows the Brier skill scores (BSSs) for three different thresholds using CLIM as the reference method again. Positive BSSs show that the method has more predictive skill than the reference; values below zero show less skill than CLIM. For threshold 0 mm day^{-1} (precipitation yes/no), it can be

TABLE 1. Summary of all four methods used for verification in section 4b. The second and third columns indicate whether the results in the verification are spatially out of sample (OOS) and/or temporally OOS, respectively. The fourth column shows whether the method provides spatial predictions or not.

Method description Abbreviation	Spatially OOS	Temporally OOS	Spatial Prediction
Climatology: CLIM	No	yes	yes
ECMWF ensemble: ENS	No	yes	yes
Stationwise: STN	No	yes	no
Spatial SAMOS: SAMOS	yes	yes	yes

seen that the ENS performs poorly, even worse than the climatology. This is mainly caused by a wet bias in the ENS (not shown), which depends on the design of the ENS predicting an average over a relatively large grid cell. Both postprocessing methods perform significantly better than the climatology. Overall, SAMOS shows the best performance even for long forecast horizons. Figures 5b and 5c show the same verification for 1 and 10 mm day^{-1}, respectively. For these thresholds, ENS is better than CLIM but outperformed by the postprocessing methods. For large thresholds (Fig. 5c) and large forecast horizons, all methods become very similar. Differences between them are no longer significant.

As last measure of performance, verification rank histograms and probability integral transform (PIT) histograms are shown in Fig. 6 for the ENS and SAMOS 1-day-ahead and 6-day-ahead forecasts to assess the calibration (Gneiting et al. 2007). In general, a more uniformly distributed histogram shows better calibration. A concave shape indicates that the forecasted distribution is too tight (underdispersive); a convex shape indicates that the distribution is too wide (overdispersive).

The verification rank histogram assesses the calibration of discrete distributions as provided by the 50 + 1 members of the ENS, yielding to 52 possible ranks. For each pair of total precipitation forecasts and observations, the rank is evaluated. Observations falling below the lowest ensemble member forecast are assigned to rank 1; observations falling above the highest ensemble member forecast are assigned to rank 52. All others are assigned to the ranks 2–51 with respect to the ensemble distribution as shown in Figs. 6a and 6c. The pronounced concave shape of the rank histogram indicates a strong underdispersion of the raw ENS such that a large fraction falls into the tails of the distribution or even outside.

The PIT histogram shows a similar measure for probabilistic forecasts. For each observation/forecast pair, the quantile conditional on the observed value is evaluated ([0.0 − 1.0]) and pooled into equidistant bins. For easy comparison with the rank histogram, we have chosen 52 uniformly distributed bins as shown in Figs. 6b and 6d. SAMOS is much better calibrated than the ENS, but the convex shape indicates that the distribution of the SAMOS is slightly wider than what is observed (overdispersive).

5. Discussion and conclusions

In this study, the standardized anomaly model output statistics (SAMOS) model has been extended and

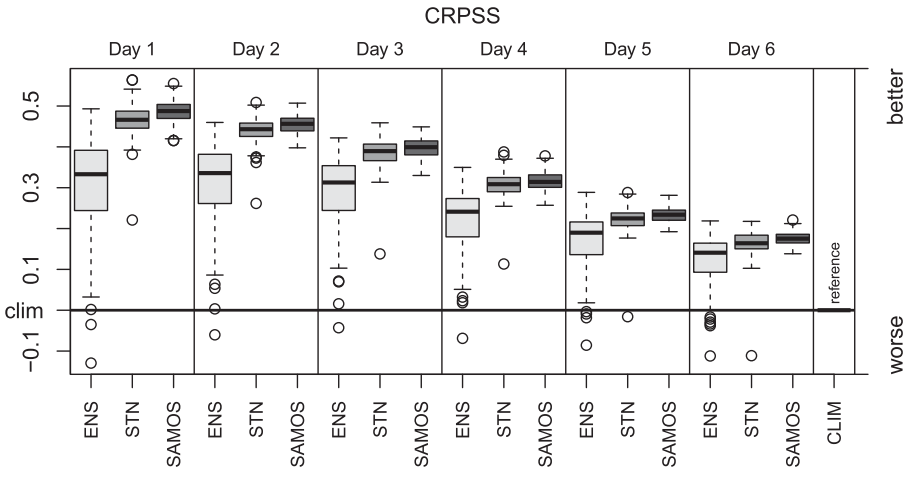

FIG. 4. CRPSS with climatology from Eq. (6) as reference. (from left to right) The boxes show the model performance for 1-day-ahead to 6-day-ahead forecasts. Each box contains three box-and-whisker plots for the (left) raw ENS, and the two postprocessing methods (middle) STN and (right) SAMOS. Each one contains 117 stationwise-mean skill scores. The boxes show the upper and lower quartile, and the whiskers show the 1.5 interquartile range. Additionally, the median (black bar) and the outliers (circles) are plotted. Values below 0 indicate stations with less skill than the climatology. The higher the values, the better the performance of the method.

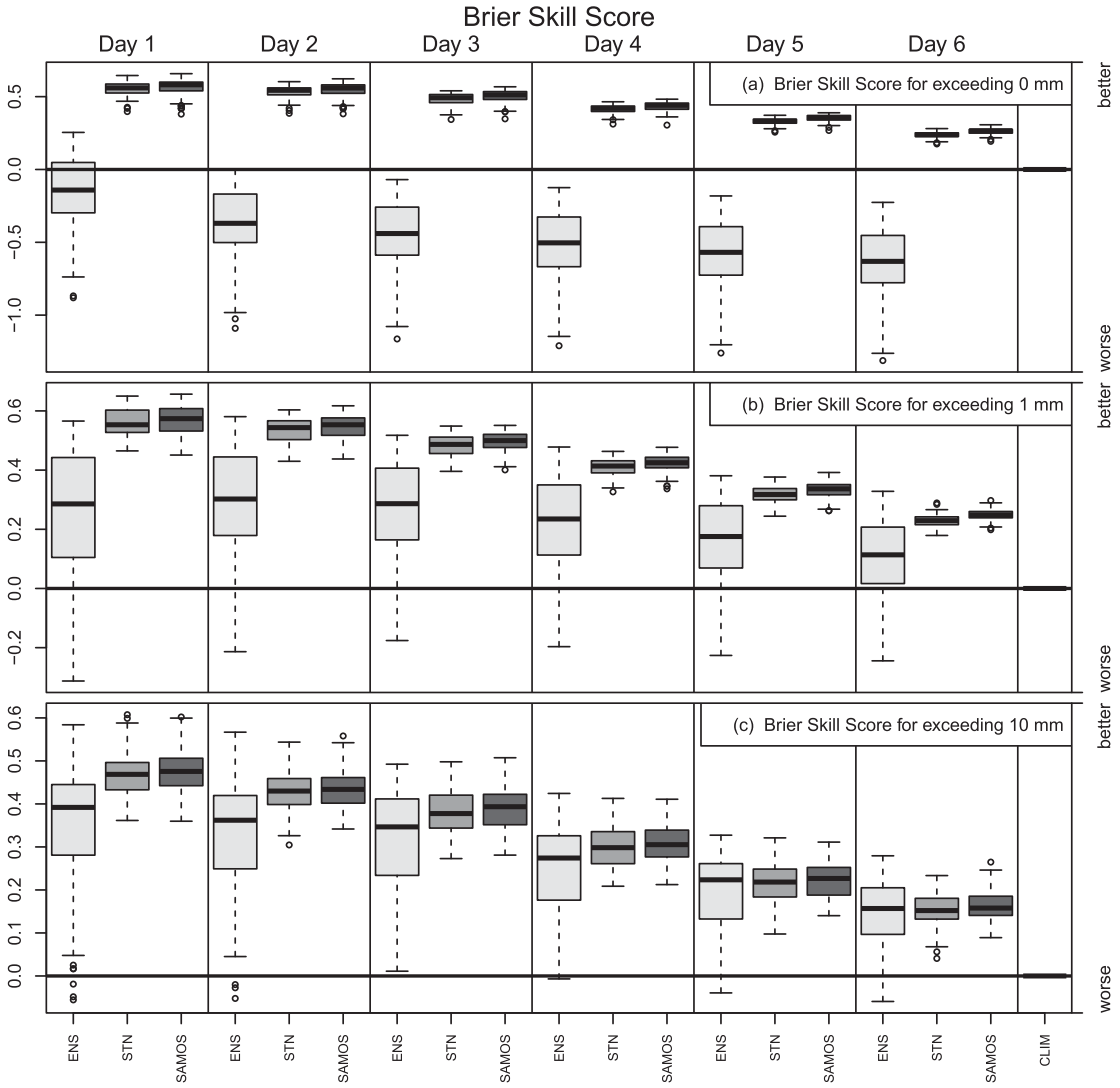

FIG. 5. BSSs for three different thresholds using climatology from Eq. (6) as reference: (a) 0, (b) 1, and (c) 10 mm day^{-1}. The specifications of the box-and-whisker plots are as in Fig. 4. The frequency of the daily total precipitation is used for ENS, whereas the probabilities for the two postprocessing methods STN and SAMOS are derived from the predicted distribution. (from left to right) Scores for 1-day-ahead to 6-day-ahead forecasts. The higher the values, the better the performance of the method.

applied to daily precipitation sums. It has been shown that the concept of using standardized anomalies (Scheuerer and Büermann 2014; Dabernig et al. 2017) can be used to correct precipitation forecasts of numerical ensemble forecast models. The SAMOS postprocessing method is able to create accurate spatial predictions of daily precipitation sums over complex terrain. SAMOS uses high-resolution spatial climatologies as background information to transform the data (observations and ensemble forecasts) into standardized anomalies. This (i) removes location-dependent climatological features from the data and (ii) brings all data to a comparable level to account for the small-scale features in the study area, which are not yet resolved by the ensemble model. SAMOS returns fully probabilistic predictions for any arbitrary location within the study area, even for regions without observational sites.

To create the standardized anomalies, daily estimates of the climatological mean (location $\mu_{\bullet,\text{clim}}$) and variability (scale $\sigma_{\bullet,\text{clim}}$) are required. The observed climatology estimate is based on the method presented by Stauffer et al. (2017) using a censored logistic rather than censored Gaussian response distribution. The censored logistic distribution has been chosen for this study as the spatiotemporal climatology showed slightly better calibration. Both distributions are very similar except that the logistic distribution has somewhat heavier tails, which is partly compensated by the

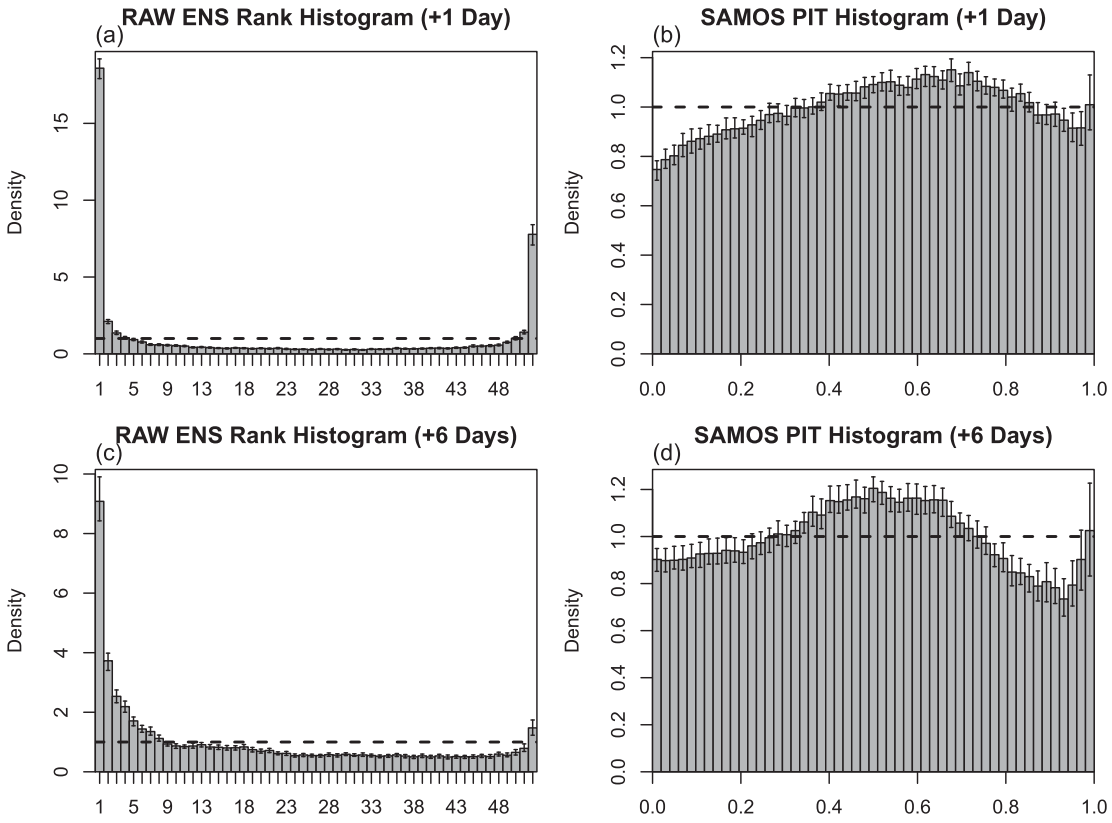

FIG. 6. (a),(c) Rank histograms of daily total precipitation sums of the raw ensemble and (b),(d) PIT histograms of the SAMOS forecasts for (top) 1-day-ahead forecasts and (bottom) 6-day-ahead forecasts. The error bars show the 95% confidence intervals of a 100× daywise random bootstrap. Rank histogram: 52 ranks (50 + 1 ensemble members). The *concave* shape indicates underdispersion. To have a similar look, the PIT histogram shows 52 bins, each of width 1/52 [first bin: (0/52 − 1/52]; second bin: (1/52 − 2/52]; and so on]. The *convex* shape indicates slight overdispersion.

additional power transformation. Overall (not shown), the predictive skill of the SAMOS using either a censored logistic distribution or a censored Gaussian distribution is very similar. The climatology of the ECMWF ensemble model is provided by the ECMWF reforecast dataset [Eq. (7)] with one reforecast run per week consisting of 4 + 1 members and covering the past 18–20 years.

Once both climatologies are known, the observations and the ensemble forecasts can be converted into standardized anomalies such that all data follow a standard logistic distribution. As all location-dependent characteristics are removed, this allows us to apply one simple regression model including all data at once. Since SAMOS uses the empirical mean and standard deviation of the standardized anomalies for training, which are based on the reforecasts, these first- and second-order moments are based on 4 + 1 members only (Roulin and Vannitsem 2012). Because of this small sample, the estimates are less precise than on current reforecasts runs, which provide 10 + 1 different ensemble members. The effect of having a larger reforecast ensemble could not be tested because of lack of overlapping data (section 2b).

The results show that the spatial SAMOS outperforms the STNs even if the SAMOS predictions are (unlike STNs) spatially out of sample. This is mainly related to the training dataset. While the STN only includes interpolated forecasts of one location, the SAMOS training dataset includes the data of all stations, leading to more robust estimates. The SAMOS calibration indicates that the assumed response distribution is not optimal. A different distribution might improve the skill and remove the need of the power transformation (Scheuerer 2014; Hamill et al. 2015).

The goal of this study is to use the SAMOS approach proposed by (Dabernig et al. 2017) and to extend the method for the application of precipitation sums or censored responses in general. While only focusing on daily precipitation sums up to day 6 in this study, it would be worthwhile to extend the forecast horizon and the study area but also to include additional covariates and to apply the SAMOS approach to other meteorological

parameters. A further SAMOS extension to account for spatiotemporal correlation structures would be of great interest. Because of the standardization, SAMOS corrects for a possible underprediction or overprediction of the ensemble over long time scales but not on a single event as only the spatial correlation structure of the EPS is considered at this stage.

As the estimation of the SAMOS requires only little computational time, the SAMOS can easily be refitted as soon as a new reforecast run is available. This ensures that the SAMOS automatically adapts itself to the latest ECMWF ensemble model version within a very short transition period. Nowadays, the ECMWF reforecast (ECMWF 2016) is run twice a week, providing 10 + 1 members, which could further improve the performance of the SAMOS but could not have been tested.

Acknowledgments. This research is part of an ongoing project funded by the Austrian Science Fund (FWF), Grant TRP 290. The computational results presented have been achieved in part using the Vienna Scientific Cluster (VSC). The observation dataset was provided by the Tyrol hydrographical service (http://ehyd.gv.at/).

APPENDIX A

Properties of the Power-Transformed Left-Censored Logistic Distribution

The probability density function λ and the cumulative distribution function Λ of a noncensored logistic distribution are defined as

$$\lambda(x \mid \mu, \sigma) = \frac{\exp\left(-\frac{x-\mu}{\sigma}\right)}{\sigma\left[1+\exp\left(-\frac{x-\mu}{\sigma}\right)\right]^2}, \tag{A1}$$

$$\Lambda(x \mid \mu, \sigma) = \frac{1}{1+\exp\left(-\frac{x-\mu}{\sigma}\right)}. \tag{A2}$$

The density λ_0 and distribution function Λ_0 of a zero left-censored logistic distribution including the power transformation $1/p$ can then be written as

$$\lambda_0(x_i \mid \mu_i, \sigma_i, p) = \begin{cases} 0 & \text{for} \quad x_i < 0 \\ \Lambda(0 \mid \mu_i, \sigma_i) & \text{for} \quad x_i = 0 \\ \lambda(x_i^{1/p} \mid \mu_i, \sigma_i) & \text{else} \end{cases}, \tag{A3}$$

$$\Lambda_0(x_i \mid \mu_i, \sigma_i, p) = \begin{cases} 0 & \text{for all: } x_i < 0 \\ \Lambda(x_i^{1/p} \mid \mu_i, \sigma_i) & \text{else} \end{cases}, \tag{A4}$$

where both are set to zero below the censoring point at 0. For $x_i^{1/p} \geq 0$, both follow the density and distribution function of the noncensored logistic distribution (λ and Λ respectively) except that the density $\lambda_0(x_i = 0 \mid \mu, \sigma)$ depicts the point mass at the censoring point, which conforms the distribution function evaluated at 0. This also directly specifies the probability of precipitation $P(x_i > 0)$ defined as the probability that precipitation will be observed at a certain location/time:

$$P(x_i > 0 \mid \mu_i, \sigma_i) = 1 - P(x_i \leq 0) = 1 - \Lambda(0 \mid \mu_i, \sigma_i). \tag{A5}$$

The probability of exceeding a certain threshold can be derived for any threshold $\kappa \geq 0$:

$$P(x_i > \kappa \mid \mu_i, \sigma_i, p) = 1 - P(x_i \leq \kappa) = 1 - \Lambda(\kappa^{1/p} \mid \mu_i, \sigma_i). \tag{A6}$$

Furthermore, the expectation of the distribution on the original scale has to be evaluated. The expectation on the original scale x in millimeters per day can be retrieved using

$$E[x \mid \mu_i, \sigma_i, p] = \int_0^\infty x\lambda(x^{1/p} \mid \mu_i, \sigma_i)\frac{x^{[(1/p)-1]}}{p}\, dx. \tag{A7}$$

A last property of interest is the median of the distribution, again on the original scale x in mm day^{-1}. Parameter μ in Eqs. (1) and (4) describes the latent unobservable location. The median is then given as

$$\text{median}(x \mid \mu, p)\begin{cases} 0 & \text{for} \quad \mu \leq 0 \\ \mu^p & \text{else} \end{cases}. \tag{A8}$$

APPENDIX B

Error Measures Used for Verification

As a fully probabilistic score, the CRPS is shown in the verification section of this article. The mean CRPS of a zero left-censored power-transformed logistic distribution (see appendix A) can be written as

$$\text{CRPS} = \frac{1}{N}\sum_{i=1}^{N}\int_0^\infty [\Lambda_0(x^{1/p} \mid \mu_i, \sigma_i) - H(x)]^2\, dx, \tag{B1}$$

where x is the response variable on the original scale in millimeters per day, N is the number of forecasts included, Λ_0 is the CDF of the forecasted distribution [Eq. (A4)], and H is the CDF of the observation represented by a Heaviside step function, which takes 0 for all $x <$ observations and 1 otherwise. While x is on the original scale (mm day^{-1}), both distributional

parameters, location μ and scale σ, are on the power-transformed scale. Therefore, the power transformation $1/p$ is required to evaluate the CDF Λ_0. As no analytic solution has been found, the CRPS is evaluated by quantile sampling with $n = 2000$.

The CRPS is shown as a skill score (CRPSS) in this article. A skill score shows the performance against a reference method. As the CRPS can only take non-negative values, the CRPSS can be written as

$$\mathrm{CRPSS} = 1 - \frac{\mathrm{CRPS}}{\mathrm{CRPS}_{\mathrm{ref}}}, \tag{B2}$$

where the CRPS of the method to test is in the numerator, and the CRPS of the reference method is in the denominator. Values below 0 indicate that the tested method performs worse than the reference. CPRSS values can take values in the range of $[-\infty, 1]$.

As a second measure, the BSS is shown to verify the skill of the forecast probabilities. One of the most frequently used thresholds for precipitation forecasts is 0 mm, also known as the probability of precipitation. As an EPS does not provide fully probabilistic forecasts, the frequency of the daily total precipitation sum is used as an estimator of the probability. As an example, if half of all ensemble members predict no precipitation, the other half does; the frequency is 0.5 and can be seen as a probability of ~50% if the number of ensemble members is sufficiently large. The Brier score can then be written as

$$\mathrm{BS}(\kappa) = \frac{1}{N}\sum_{i=1}^{N}\left[P(x_i > \kappa \mid \mu_i, \sigma_i, p) - o_i(t)\right]^2, \tag{B3}$$

where N is again the number of forecasts included, P_i the predicted probability that an event exceeds threshold κ [Eq. (A6)], and o_i the binary observation which takes 0 for all observations $x_i \leq \kappa$ and 1 otherwise. Correspondingly, the Brier skill score is defined as

$$\mathrm{BSS} = 1 - \frac{\mathrm{BSS}}{\mathrm{BSS}_{\mathrm{ref}}}. \tag{B4}$$

REFERENCES

Ben Bouallègue, Z., and S. E. Theis, 2014: Spatial techniques applied to precipitation ensemble forecasts: From verification results to probabilistic products. *Meteor. Appl.*, **21**, 922–929, doi:10.1002/met.1435.

Box, G. E. P., and D. R. Cox, 1964: An analysis of transformations. *J. Roy. Stat. Soc.*, **26B**, 211–252.

Buizza, R., P. L. Houtekamer, G. Pellerin, Z. Toth, Y. Zhu, and M. Wei, 2005: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Wea. Rev.*, **133**, 1076–1097, doi:10.1175/MWR2905.1.

Bundesministerium für Land und Forstwirtschaft, Umwelt und Wasserwirtschaft, 2016: Abteilung IV/4—Wasserhaushalt. Accessed 29 February 2016. [Available online at http://ehyd.gv.at.]

Dabernig, M., G. J. Mayr, J. W. Messner, and A. Zeileis, 2017: Spatial ensemble post-processing with standardized anomalies. *Quart. J. Roy. Meteor. Soc.*, doi:10.1002/qj.2975, in press.

ECMWF, 2016: Re-forecast for medium and extended forecast range. ECMWF, accessed 9 June 2016. [Available online at http://www.ecmwf.int/en/forecasts/documentation-and-support/re-forecast-medium-and-extended-forecast-range.]

Epstein, E. S., 1969: Stochastic dynamic prediction. *Tellus*, **21**, 739–759, doi:10.1111/j.2153-3490.1969.tb00483.x.

Fraley, C., A. E. Raftery, and T. Gneiting, 2010: Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging. *Mon. Wea. Rev.*, **138**, 190–202, doi:10.1175/2009MWR3046.1.

Frei, C., and C. Schär, 1998: A precipitation climatology of the Alps from high-resolution rain-gauge observations. *Int. J. Climatol.*, **18**, 873–900, doi:10.1002/(SICI)1097-0088(19980630)18:8<873::AID-JOC255>3.0.CO;2-9.

Gebetsberger, M., J. W. Messner, G. J. Mayr, and A. Zeileis, 2016: Tricks for improving non-homogeneous regression for probabilistic precipitation forecasts: Perfect predictions, heavy tails, and link functions. University of Innsbruck Working Papers in Economics and Statistics 2016-28, 25 pp. [Available online at http://EconPapers.repec.org/RePEc:inn:wpaper:2016-28.]

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098–1118, doi:10.1175/MWR2904.1.

——, F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc.*, **69B**, 243–268, doi:10.1111/j.1467-9868.2007.00587.x.

Hagedorn, R., R. Buizza, T. M. Hamill, M. Leutbecher, and T. N. Palmer, 2012: Comparing TIGGE multimodel forecasts with reforecast-calibrated ECMWF ensemble forecasts. *Quart. J. Roy. Meteor. Soc.*, **138**, 1814–1827, doi:10.1002/qj.1895.

Hamill, T. M., 2012: Verification of TIGGE multimodel and ECMWF reforecast-calibrated probabilistic precipitation forecasts over the contiguous United States. *Mon. Wea. Rev.*, **140**, 2232–2252, doi:10.1175/MWR-D-11-00220.1.

——, J. S. Whitaker, and S. L. Mullen, 2006: Reforecasts: An important dataset for improving weather predictions. *Bull. Amer. Meteor. Soc.*, **87**, 33–46, doi:10.1175/BAMS-87-1-33.

——, R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Mon. Wea. Rev.*, **136**, 2620–2632, doi:10.1175/2007MWR2411.1.

——, M. Scheuerer, and G. T. Bates, 2015: Analog probabilistic precipitation forecasts using GEFS reforecasts and climatology-calibrated precipitation analyses. *Mon. Wea. Rev.*, **143**, 3300–3309, doi:10.1175/MWR-D-15-0004.1.

Hutchinson, M. F., 1998: Interpolation of rainfall data with thin plate smoothing splines—Part I: Two dimensional smoothing of data with short range correlation. *J. Geogr. Inf. Decis. Anal.*, **2**, 168–185.

Isotta, F. A., and Coauthors, 2014: The climate of daily precipitation in the Alps: Development and analysis of a high-resolution grid dataset from pan-Alpine rain-gauge data. *Int. J. Climatol.*, **34**, 1657–1675, doi:10.1002/joc.3794.

Jarvis, A., H. I. Reuter, A. Nelson, and E. Guevara, 2008: SRTM 90m digital elevation database, version 4.1. Consultative Group on International Agricultural Research Consortium for Spatial Information, accessed 29 February 2016. [Available online at http://srtm.csi.cgiar.org.]

Kaiser, M., and Coauthors, 2014: Statistisches Handbuch Bundesland Tirol 2014. Land Tirol Rep., 422 pp. [Available online at https://www.tirol.gv.at/fileadmin/themen/statistik-budget/statistik/downloads/Statistisches_Handbuch_2014.pdf.]

Lerch, S., and T. L. Thorarinsdottir, 2013: Comparison of non-homogeneous regression models for probabilistic wind speed forecasting. *Tellus*, **65**, 21206, doi:10.3402/tellusa.v65i0.21206.

Messner, J. W., G. J. Mayr, D. S. Wilks, and A. Zeileis, 2014a: Extending extended logistic regression: Extended versus separate versus ordered versus censored. *Mon. Wea. Rev.*, **142**, 3003–3014, doi:10.1175/MWR-D-13-00355.1.

——, ——, A. Zeileis, and D. S. Wilks, 2014b: Heteroscedastic extended logistic regression for postprocessing of ensemble guidance. *Mon. Wea. Rev.*, **142**, 448–456, doi:10.1175/MWR-D-13-00271.1.

——, ——, and ——, 2016: Heteroscedastic censored and truncated regression with crch. *R J.*, **8**, 173–181. [Available online at https://journal.r-project.org/archive/2016-1/messner-mayr-zeileis.pdf.]

Mullen, S. L., and R. Buizza, 2001: Quantitative precipitation forecasts over the United States by the ECMWF ensemble prediction system. *Mon. Wea. Rev.*, **129**, 638–663, doi:10.1175/1520-0493(2001)129<0638:QPFOTU>2.0.CO;2.

Roulin, E., and S. Vannitsem, 2012: Postprocessing of ensemble precipitation predictions with extended logistic regression based on hindcasts. *Mon. Wea. Rev.*, **140**, 874–888, doi:10.1175/MWR-D-11-00062.1.

Roulston, M. S., and L. A. Smith, 2003: Combining dynamical and statistical ensembles. *Tellus*, **55**, 16–30, doi:10.1034/j.1600-0870.2003.201378.x.

Scheuerer, M., 2014: Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quart. J. Roy. Meteor. Soc.*, **140**, 1086–1096, doi:10.1002/qj.2183.

——, and L. Büermann, 2014: Spatially adaptive post-processing of ensemble forecasts for temperature. *J. Roy. Stat. Soc.*, **63C**, 405–422, doi:10.1111/rssc.12040.

——, and T. M. Hamill, 2015: Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Mon. Wea. Rev.*, **143**, 4578–4596, doi:10.1175/MWR-D-15-0061.1.

Sloughter, J. M. L., A. E. Raftery, T. Gneiting, and C. Fraley, 2007: Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Mon. Wea. Rev.*, **135**, 3209–3220, doi:10.1175/MWR3441.1.

Statistik Austria, 2016: Bevölkerung. Accessed 22 June 2016. [Available online at https://www.statistik.at/web_de/statistiken/menschen_und_gesellschaft/bevoelkerung/index.html.]

Stauffer, R., G. J. Mayr, J. W. Messner, N. Umlauf, and A. Zeileis, 2017: Spatio-temporal precipitation climatology over complex terrain using a censored additive regression model. *Int. J. Climatol.*, doi:10.1002/joc.4913, in press.

Stidd, C. K., 1973: Estimating the precipitation climate. *Water Resour. Res.*, **9**, 1235–1241, doi:10.1029/WR009i005p01235.

Thorarinsdottir, T. L., and T. Gneiting, 2010: Probabilistic forecasts of wind speed: Ensemble model output statistics by using heteroscedastic censored regression. *J. Roy. Stat. Soc.*, **173A**, 371–388, doi:10.1111/j.1467-985X.2009.00616.x.

Wilks, D. S., 2009: Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteor. Appl.*, **16**, 361–368, doi:10.1002/met.134.

Article IV

Stauffer, R., Mayr, G.J., Messner J.W., and Zeileis A. (2018). *Hourly Probabilistic Snow Forecasts over Complex Terrain: A Hybrid Ensemble Postprocessing Approach*. Advances in Statistical Climatoloy, Meteorology and Oceanography, 4, 65-86, doi:10.5194/ASCMO-4-65-2018.

Recent peer-reviewed journal on the intersection of atomspheric science and statistics (published by Copernicus), not yet listed in JCR.

Adv. Stat. Clim. Meteorol. Oceanogr., 4, 65–86, 2018
https://doi.org/10.5194/ascmo-4-65-2018

ASCMO
Open Access

Hourly probabilistic snow forecasts over complex terrain: a hybrid ensemble postprocessing approach

Reto Stauffer[1], Georg J. Mayr[2], Jakob W. Messner[3], and Achim Zeileis[1]

[1]Department of Statistics, Faculty of Economics and Statistics, Universität Innsbruck, Universitätsstraße 15, 6020 Innsbruck, Austria
[2]Institute of Atmospheric and Cryospheric Sciences, Faculty of Geo- and Atmospheric Sciences, Universität Innsbruck, Innrain 52, 6020 Innsbruck, Austria
[3]Department of Electrical Engineering, Technical University of Denmark, Elektrovej, Building 325, 2800 Kgs. Lyngby, Denmark

Correspondence: Reto Stauffer (reto.stauffer@uibk.ac.at)

Received: 27 March 2018 – Revised: 24 August 2018 – Accepted: 12 October 2018 – Published: 14 December 2018

Abstract. Accurate and high-resolution snowfall and fresh snow forecasts are important for a range of economic sectors as well as for the safety of people and infrastructure, especially in mountainous regions. In this article a new hybrid statistical postprocessing method is proposed, which combines standardized anomaly model output statistics (SAMOS) with ensemble copula coupling (ECC) and a novel re-weighting scheme to produce spatially and temporally high-resolution probabilistic snow forecasts. Ensemble forecasts and hindcasts of the European Centre for Medium-Range Weather Forecasts (ECMWF) serve as input for the statistical postprocessing method, while measurements from two different networks provide the required observations.

This new approach is applied to a region with very complex topography in the eastern European Alps. The results demonstrate that the new hybrid method allows one not only to provide reliable high-resolution forecasts, but also to combine different data sources with different temporal resolutions to create hourly probabilistic and physically consistent predictions.

1 Introduction

Large parts of our daily social and economic life strongly rely on weather forecasts. In this article we focus on the governmental area of Tyrol, Austria, which is located in the eastern Alps and consists of a large number of narrow valleys surrounded by high mountains. The economic backbone of Tyrol is tourism with more than 5.3 million visitors and more than 25 million overnight stays recorded during the winter season 2013/14 (Amt der Tiroler Landesregierung, 2014). In winter tourism strongly focuses on Alpine outdoor sports such as skiing and back-country skiing, for which resorts and skiing areas need sufficient amounts of snow and good snow conditions. On the other hand, the "white gold" can also cause hazardous situations. During the winter seasons 2009–2016 145 people died in avalanche accidents in Austria (Lawinenwarndienst Tirol, 2009–2017), of which more than half of all events and deaths occurred in Tyrol. Furthermore, severe snow events can obstruct traffic on roads, on train tracks and at airports. Accurate and reliable forecasts of fresh snow and snowfall for the region of Tyrol are therefore of high importance for the public and also for decision makers or warning services (see, e.g., Zhu et al., 2002; Palmer, 2002; Neal et al., 2014; Knox et al., 2015; Raftery, 2016).

Weather forecasts are typically provided by numerical weather prediction (NWP) models predicting the future atmospheric state on a global or regional scale. Due to different influencing factors such as the model resolution, necessary approximations and parameterizations but also imperfect initial conditions and the chaotic behavior of the atmosphere, these forecasts are never fully exact. Ensemble prediction systems (EPSs) address these issues by running sev-

eral independent forecasts for the same day using different and slightly perturbed initial conditions and model formulations to provide valuable additional information about the uncertainty of a specific weather forecast. Due to the spatial discretization of the underlying NWP model the EPS can only depict information on a grid-scale level and is not able to provide reliable information on the point scale. Thus, EPS forecasts typically show too little spread (Hagedorn et al., 2012; Mullen and Buizza, 2001) and require additional correction of the EPS uncertainty to enhance the predictive skill for specific locations. One widely accepted procedure to reduce possible forecast errors and to adjust the uncertainty information is statistical ensemble postprocessing. Statistical postprocessing methods use historical weather forecasts and the corresponding observations to detect and correct possible systematic EPS errors.

A wide range of different ensemble postprocessing methods have been proposed, including analog methods (Hamill et al., 2006, 2015), ensemble dressing methods (Roulston and Smith, 2003), extended logistic regression (Wilks, 2009; Bouallègue and Theis, 2014; Messner et al., 2014b), a non-homogeneous mixture model approach with similarities to Bayesian model averaging (BMA; Sloughter et al., 2007; Fraley et al., 2010), or distributional regression methods. Distributional regression models optimize the parameters of a pre-specified response distribution to correct for both errors in the mean and errors in the uncertainty, given a set of covariates. One of the earliest and most well-known approaches is the ensemble model output statistics (EMOS) approach first published by Gneiting et al. (2005) and applied to near-surface temperature. This approach has further been extended by Thorarinsdottir and Gneiting (2010), Lerch and Thorarinsdottir (2013), Scheuerer (2014), Scheuerer and Hamill (2015), Messner et al. (2014a), Scheuerer (2014), Scheuerer and Hamill (2015) and many others for different meteorological quantities using different response distributions and optimization approaches.

Originally, distributional regression was only applied to specific locations, but has also been extended for spatial and even spatio-temporal corrections of the ensemble forecasts. Many of these extensions are based on anomalies (Scheuerer and Büermann, 2014) or standardized anomalies (Dabernig et al., 2017; Stauffer et al., 2017b) to account for location-specific characteristics in mean and variance and create corrected and fully probabilistic spatial predictions of temperature and daily precipitation sums over potentially complex terrain.

In terms of snow prediction several difficulties have to be considered. The availability and quality of good and reliable snow observations are sparse, even in the region of Tyrol. Measuring snow can be tricky due to possible snow drift, melting processes, or liquid water input (rain) between two observation times, which can yield large measurement errors (Rasmussen et al., 2012). Overall, the amount and quality of snow measurements make it very difficult to train reliable spatial postprocessing models.

An alternative approach to predict fresh snow amounts is to make use of precipitation and temperature forecasts rather than directly to predict snow. The postprocessed temperature and precipitation forecasts can then be used as a proxy to retrieve fresh snow amounts and snowfall forecasts. The temperature forecasts are on the one hand required to determine whether precipitation reaches the ground as rain or snow and on the other hand to estimate the snow density. Snow density and its alteration are affected by the prevalence of inversions, additional cooling effects due to melting and evaporation of hydrometeors, and other local effects, and are thus an extremely complex issue itself. For simplicity we will only regard the problem of whether precipitation occurs as snow or rain and assume that precipitation will fall as snow as soon as the 2 m dry air temperature falls below $+1.2\,^{\circ}\mathrm{C}$, a threshold used in the literature for the European Alps (Rohregger, 2008; Bellaire et al., 2011).

Major challenges of converting probabilistic precipitation and temperature forecasts into fresh snow predictions are the very different temporal resolutions of ensemble predictions, temperature observations, and precipitation observations. European Centre for Medium-Range Weather Forecast (ECMWF) hindcast and EPS forecasts, which we use in this study, have a temporal resolution of 6 and 1 h, respectively, temperature observations are usually available hourly, and precipitation or snow heights are often only measured once or a few times a day.

In this article we propose a new hybrid approach that combines standardized anomaly model output statistics (SAMOS; Dabernig et al., 2017; Stauffer et al., 2017b) with ensemble copula coupling (ECC; Schefzik et al., 2013) and a novel re-weighting scheme to combine these data to

i. create full probabilistic spatial predictions,

ii. provide probabilistic temperature and precipitation forecasts on an hourly temporal scale, and

iii. create a physically consistent copula (pair of temperature and precipitation) which can be used to

iv. create spatially and temporally high-resolution snowfall and fresh snow amount forecasts.

The structure of this article is as follows. Section 2 introduces the different statistical methods required to achieve the desired goal. The methods section is followed by the description of the different data sets used in this study (Sect. 3) and the explicit specification of the statistical models (Sect. 4) used in the results section (Sect. 5). At the end the results and limitations of this approach will be discussed (Sect. 6).

2 Methods

This section contains the three methodological blocks required to create probabilistic snow forecasts. Distributional regression is explained in Sect. 2.1 followed by the required extensions for the SAMOS in Sect. 2.2. Section 2.3 shows the ensemble copula coupling (ECC) approach to generate a postprocessed ensemble followed by the re-weighting procedure in Sect. 2.4 which is required to transform daily precipitation sums into hourly predictions. Finally, hourly temperature and precipitation sums will be converted into probabilities of snowfall and fresh snow amounts in Sect. 2.5.

2.1 Distributional regression

Statistical methods considering all parameters of a specific response distribution can be summarized as "distributional regression models". The EMOS for temperature using a normal response distribution as originally suggested by Gneiting et al. (2005) can be seen as a classical example of this family.

Imagine a time series of 2 m temperature observations $\boldsymbol{y} = \{y_i\}_{i=1,\ldots,N}$ at a specific site and the corresponding ensemble forecasts of the 2 m temperature from the EPS $\mathbf{x} = \{x_{\mathrm{im}}\}_{i=1,\ldots,N}^{m=1,\ldots,M}$ where N denotes the total sample size of the data set and M the number of ensemble members. x_{im} is the individual 2 m temperature prediction of the NWP for date/time i of member m. The EMOS, which slightly differs from the original EMOS as proposed by Gneiting et al. (2005), is specified as

$$
\begin{aligned}
y_i &\sim \mathcal{N}(\mu_i, \sigma_i), && (1)\\
\mu_i &= \beta_0 + \beta_1 \cdot \overline{\boldsymbol{x}}_i, && (2)\\
\log(\sigma_i) &= \gamma_0 + \gamma_1 \cdot \langle \boldsymbol{x}_i \rangle. && (3)
\end{aligned}
$$

The response y_i is assumed to follow a normal distribution $\mathcal{N}$ with the two distributional parameters μ_i (location or mean) and σ_i (scale or standard deviation). Both parameters are expressed by a linear predictor including an intercept (β_0/γ_0) and a slope coefficient (β_1/γ_1) for a covariate. While the location μ_i depends on the ensemble mean $\overline{\boldsymbol{x}}_i$ over all members $m = 1, \ldots, M$ for each individual sample i, the log scale depends on the logarithm of the corresponding ensemble standard deviation denoted as $\langle \boldsymbol{x}_i \rangle$. The log link on σ_i ensures positive variance in predictions.

The coefficients $\boldsymbol{\theta} = (\beta_0, \beta_1, \gamma_0, \gamma_1)$ can be estimated by using an appropriate M estimator such as the maximum-likelihood estimator maximizing the likelihood:

$$
\hat{\boldsymbol{\theta}} = \underset{\theta}{\operatorname{argmax}} \left(\prod_{i=1}^{N} \phi\left(\frac{y_i - \mu_i}{\sigma_i} \right) \right), \qquad (4)
$$

where $\phi\left(\frac{y_i - \mu_i}{\sigma_i}\right)$ denotes the standard normal probability density function (PDF) evaluated at each individual $i = 1, \ldots, N$ in the data set.

For the daily precipitation sums the model shown in Eqs. (1)–(3) can be improved by replacing the response distribution and adding an additional covariate z which allows one to account for EPS forecasts where the majority of all EPS members predict no precipitation. Following the work of Gebetsberger et al. (2017) and Stauffer et al. (2017a), the model specification can be written as follows:

$$
\begin{aligned}
y_i^{1/p} &= \mathcal{L}_0(\mu_i, \sigma_i), && (5)\\
\mu_i &= \beta_0 + \beta_1 \cdot \overline{\boldsymbol{x}_i^{1/p}} \cdot (1 - z_i) + \beta_2 \cdot z_i, && (6)\\
\log(\sigma_i) &= \gamma_0 + \gamma_1 \cdot \langle \boldsymbol{x}_i^{1/p} \rangle \cdot (1 - z_i). && (7)
\end{aligned}
$$

The power-transformed observations y_i are assumed to follow a left-censored logistic distribution $\mathcal{L}_0$ censored at 0 and a power parameter $p = 1.35$. The additional covariate z_i takes 1 if 80 % or more of all ensemble members predict less than 0.05 mm over 24 h and 0 otherwise and is used to handle unanimous predictions (cf. Gebetsberger et al., 2017). The corresponding M estimator can be written as

$$
\hat{\boldsymbol{\theta}} = \underset{\theta}{\operatorname{argmax}} \left(\prod_{i=1}^{N} f\left(\frac{y_i^{1/p} - \mu_i}{\sigma_i} \right) \right)
$$

$$
\text{with } f = \begin{cases} \Lambda\left(\frac{-\mu_i}{\sigma_i}\right) & \text{if } y_i = 0 \\ \lambda\left(\frac{y_i^{1/p} - \mu_i}{\sigma_i}\right) & \text{else} \end{cases}, \qquad (8)
$$

where λ is the PDF and Λ the cumulative distribution function (CDF) of the standard logistic distribution.

2.2 SAMOS

While the model specifications in Eqs. (1)–(3) and (5)–(7) work well for single stations, an extension is required for spatial and/or spatio-temporal ensemble postprocessing. In the following, we will employ the SAMOS approach (Dabernig et al., 2017; Stauffer et al., 2017b) for this purpose. Its basic idea is to remove location- and time-specific characteristics from the observation and EPS data by transforming them into standardized anomalies. This transformation then allows one to fit a single postprocessing model that is valid for the whole area and all season and can thus be applied to any location and time.

Standardized anomalies of the observations ($\boldsymbol{y}^*$) and EPS forecasts ($\mathbf{x}^*$, for each member $m \in M$) will be characterized by a superscript asterisk from here on and are defined as

$$
y_i^* = \frac{y_i - \widetilde{\mu}_{y,i}}{\widetilde{\sigma}_{y,i}} \quad \text{and} \quad x_{\mathrm{im}}^* = \frac{x_{\mathrm{im}} - \widetilde{\mu}_{x,i}}{\widetilde{\sigma}_{x,i}}. \qquad (9)
$$

$\widetilde{\mu}_{\bullet,i}$ and $\widetilde{\sigma}_{\bullet,i}$ are the estimates of the climatological location and scale for each required quantity and depend on the location and season respectively. A comprehensive description of how these climatologies are specified can be found

in Appendix A. Once the climatologies and thus the standardized anomalies are known, the SAMOS regression coefficients can be estimated using Eqs. (1)–(8) by simply replacing y_i and $\boldsymbol{x}_i$ by their corresponding standardized anomalies y_i^* and $\boldsymbol{x}_i^*$ (except in the condition in Eq. (8) where $y_i = 0$ is not replaced). Given a new EPS forecast, the postprocessed predictions can be obtained by applying the SAMOS correction. As the regression coefficients $\hat{\boldsymbol{\theta}}$ are time and location independent, the correction can be performed on the EPS grid scale. Spatial predictions can be retrieved by bilinearly interpolating the resulting location ($\boldsymbol{\mu}^*$) and scale ($\boldsymbol{\sigma}^*$) parameters to the desired spatial resolution and transforming the results back to the original scale (e.g., °C or mm). Algorithm 1 contains the pseudo-code for the SAMOS procedure as used for this article.

Algorithm 1 SAMOS postprocessing procedure. Detailed description and graphical representation in Appendix A.

1. Compute standardized anomalies of observations:

   Input. Observations $\boldsymbol{y}$ (temperature: hourly, precipitation 24-hourly; Sect. 3.3). Each y_i is an observation at a given station and time point.

   i. Estimate spatio-temporal climatology $\tilde{\mu}_{y,i}$, $\tilde{\sigma}_{y,i}$ for response $\boldsymbol{y}$ including seasonal and station location characteristics. Separate climatologies are estimated for temperature and precipitation (Appendix A). *Note.* This allows to obtain climatological parameters $\tilde{\mu}_{y,i}$, $\tilde{\sigma}_{y,i}$ not only at observed locations/time points i but also at new locations/time points j.

   ii. Compute standardized anomalies $y_i^* = \left(y_i - \tilde{\mu}_{y,i}\right)/\tilde{\sigma}_{y,i}$ separate for temperature and precipitation (Sect. 2.2).

   Output. Spatio-temporal climatologies and standardized anomalies y_i^* of the observations at station level.

2. Calculate standardized anomalies of NWP forecasts:

   Input. Gridded hindcasts (temperature: 6-hourly, precipitation: 24-hourly; Sect. 3.2) separately for each required covariate $\boldsymbol{x} \in \mathbf{X}$. Each $\boldsymbol{x}_\mathrm{g}$ is one hindcast at a given grid point for a specific time and forecast lead time with $M = 10 + 1$ members.

   i. Estimate gridded model climatologies $\tilde{\mu}_{x,\mathrm{g}}$, $\tilde{\sigma}_{x,\mathrm{g}}$ separately for each $\boldsymbol{x} \in \mathbf{X}$ (Appendix A).

   ii. Compute standardized anomalies $x_\mathrm{gm}^* = \left(x_\mathrm{gm} - \tilde{\mu}_{x,\mathrm{g}}\right)/\tilde{\sigma}_{x,g}$ for all required covariates $\boldsymbol{x} \in \mathbf{X}$ for each $m \in M$ (Sect. 2.2).

   iii. Bilinearly interpolate standardized anomalies x_gm^* to all station locations to obtain x_im^*.

   Output. Model climatologies and standardized anomalies x_im^* at station level (temperature: 6-hourly, precipitation: 24-hourly).

3. Estimate SAMOS models:

   Inputs. Standardized anomalies (y_i^*, x_im^*) from Steps 1 and 2 (temperature: 6-hourly, precipitation 24-hourly). As y_i^* and $\boldsymbol{x}_i^*$ are no longer location and season dependent: pool all data (space and time) into one combined training data set (separately for temperature and precipitation).

   Estimate the statistical models $\boldsymbol{y}^* \sim \mathcal{D}(\boldsymbol{\mu}^*, \boldsymbol{\sigma}^*)$ to get the required regression coefficients $\hat{\boldsymbol{\theta}}$, separate models for temperature and precipitation (Sects. 2.1 and 2.2).

   Output. Two sets of estimated regression coefficients $\hat{\boldsymbol{\theta}}$ for temperature and precipitation postprocessing, respectively.

4. Predict temperature and precipitation given a new NWP forecast:

   Inputs. Gridded EPS forecasts (Sect. 3.1), observation climatologies (Step 1i) and gridded model climatologies (Step 2i).

   i. Compute standardized anomalies at grid level for covariates $\boldsymbol{x} \in \mathbf{X}$ of each member $m \in M$ of the new EPS forecast with respect to model climatology: $x_\mathrm{gm}^* = \left(x_\mathrm{gm} - \tilde{\mu}_{x,\mathrm{g}}\right)/\tilde{\sigma}_{x,\mathrm{g}}$.

   ii. Correct forecast anomalies for each $g \in G$ using the estimated coefficients $\hat{\boldsymbol{\theta}}$ from Step 3 to get μ_g^* and σ_g^*

   iii. Interpolate parameters (μ_g^*, σ_g^*) of the postprocessed standardized anomalies to obtain μ_j^* and σ_j^* where each j corresponds to a given (arbitrary) location within the study area and a specific time point and forecast lead time.

   iv. Transform corrected μ_j^* and σ_j^* to physical scale: $\hat{y}_j \sim \mathcal{D}\left(\mu_j^* \cdot \tilde{\sigma}_{y,j} + \tilde{\mu}_{y,j},\ \sigma_j^* \cdot \tilde{\sigma}_{y,j}\right) = \mathcal{D}\left(\hat{\mu}_j,\ \hat{\sigma}_j\right)$

   Output. Postprocessed full parametric predictions at each target location (temperature: hourly, precipitation: 24-hourly; Sect. 5.6).

2.3 Ensemble copula coupling

The SAMOS procedure (Sect. 2.2) provides postprocessed probabilistic predictions for 2 m temperature as well as corrected probabilistic forecasts for 24 h precipitation sums. Due to the model specification, SAMOS allows one to retrieve predictions for any arbitrary location within the area of interest (spatial prediction) and even for all forecast steps covered by the training data set (temporal predictions) with one set of regression coefficients. This allows one to create forecasts for $+30/+54/+78$ h for the 24 h precipitation sums, and hourly forecasts for 2 m temperature for the whole study area.

In order to retrieve probabilistic snowfall forecasts from the SAMOS predictions, the marginal predictive distributions of temperature and precipitation have to be combined such that correlations between them are considered. This can be achieved by using ensemble copula coupling (ECC) proposed by Schefzik et al. (2013). The basic idea is to restore the physical coupling between two or more quantities based on the rank order structure of the raw EPS. As numerical predictions are based on physically consistent prognos-

tic equations, each EPS member provides a distinct physically meaningful combination of temperature and precipitation. This property is lost during the SAMOS postprocessing since both quantities are corrected independently. However, the coupling can be restored by drawing a sample of the postprocessed predictive distributions and rearranging the sampled values in the rank order structure of the original EPS forecasts. ECC is applied to each target location individually to restore the spatial correlation structure of the EPS.

There are different ways to draw a new sample from the postprocessed distributions. It turned out (not shown) that the quantile mapping approach with equidistant probabilities (ECC-Q; Schefzik et al., 2013) yields the best and most stable results for this application, which supports the findings of Schefzik et al. (2013). For ECC-Q, a set of $M = 50 + 1$ ensemble members is drawn from the postprocessed distribution based on equidistant probabilities. In the case of the 2 m temperature SAMOS returns hourly estimates for location $\hat{\mu}_j$ and scale $\hat{\sigma}_j$ of a Gaussian distribution (Eq. 3; Algorithm 1 step 4iv). Using the inverse Gaussian CDF $\Phi^{-1}(\boldsymbol{\pi} \mid \hat{\mu}_j, \hat{\sigma}_j)$ with equidistant probabilities $\boldsymbol{\pi} = \frac{1}{M+1}, \ldots, \frac{M}{M+1}$ a new $50 + 1$-member temperature ensemble can be retrieved from the postprocessed distribution.

The very same can be done for the daily precipitation sums using the inverse distribution function of the power-transformed left-censored logistic distribution:

$$\Lambda_0^{-1}\left(\boldsymbol{\pi} \mid \hat{\mu}_j, \hat{\sigma}_j, p\right) = \max\left(0,\ \Lambda^{-1}(\boldsymbol{\pi} \mid \hat{\mu}_j, \hat{\sigma}_j)\right)^p, \qquad (10)$$

where Λ^{-1} is the inverse CDF of the uncensored logistic distribution. Due to the left-censoring at 0, some of the M quantiles can fall on the censoring point, with an increasing number of 0s with decreasing location $\hat{\mu}_j$ and vice versa. For situations where precipitation is very unlikely $\hat{\mu}_j$ might be highly negative, which yields a postprocessed ensemble where all M members predict exactly $0\,\mathrm{mm}\,24\,\mathrm{h}^{-1}$. However, there is still the problem that our two quantities are not available on the same temporal scale. To be able to restore the full EPS rank order structure on an hourly temporal resolution the postprocessed daily precipitation sums first have to be downscaled to an hourly interval.

2.4 Precipitation re-weighting

Temperature and precipitation observation data are based on two different observational networks with different temporal resolutions. The 2 m temperature observations are available hourly, while precipitation sums are only reported once a day (details in Sect. 3.3). This temporal resolution is maintained by the SAMOS postprocessing so that it also differs for the forecasts of the different quantities.

As temperature shows a clear diurnal cycle, it is crucial to know at which time of day precipitation is expected to be observed, as the timing can highly affect the precipitation phase and thus the total fresh snow amount. Therefore, the precipitation forecasts have to be temporally downscaled before they can be combined with the temperature forecasts. For this purpose, we extend ECC (Sect. 2.3) with a novel re-weighting scheme where the daily precipitation sums are allocated to the hours of the day according to the time series of the raw EPS predictions. For example, if an EPS member predicts 10 % of its daily precipitation to fall between 10:00 and 11:00, 10 % of the corresponding precipitation forecast is allocated to this hour. This allows one to downscale each of the $M = 50 + 1$ draws from the marginal precipitation to an hourly temporal resolution and to combine the hourly precipitation predictions afterwards with the respective draws from the marginal temperature distribution. Algorithm 2 shows the temporal downscaling procedure to generate hourly precipitation copulas from the postprocessed daily precipitation sums.

Algorithm 2 Re-weighting pseudo-code for temporal downscaling of probabilistic precipitation forecasts to generate a new $50 + 1$ member copula with an hourly temporal resolution from postprocessed probabilistic daily precipitation sum forecasts provided by SAMOS.

Inputs. Gridded EPS forecasts of 24 h accumulated precipitation sums and postprocessed probabilistic 24 h precipitation sums returned by SAMOS. Index j denotes a specific target location, season, and forecast lead time.

1. Bilinearely interpolate forecasted 24 h precipitation sums of each of the $50 + 1$ EPS members to target location $j \in J$ to receive $(\mathrm{tp}_{j1}, \ldots, \mathrm{tp}_{jm}, \ldots, \mathrm{tp}_{jM})$.
2. Draw a copula $(\hat{y}_{j1}, \ldots, \hat{y}_{jm}, \ldots, \hat{y}_{jM})$ of 24 h postprocessed precipitation sums using ECC-Q drawing from the full probabilistic predictive distribution $\mathcal{L}_0(\hat{\mu}_j, \hat{\sigma}_j)^{1.35}$ returned by SAMOS (Sect. 2.2).
3. Compute correction weighting factors $\boldsymbol{\omega}_j = (\hat{y}_{j1}/\mathrm{tp}_{j1}, \ldots, \hat{y}_{jm}/\mathrm{tp}_{jm}, \ldots, \hat{y}_{jM}/\mathrm{tp}_{jM})$.
4. Correct hourly EPS time series forecasts of each member using the weights $\boldsymbol{\omega}_j$ such that the sum over 24 1-hourly precipitation sums of a specific copula member m is equal to $\hat{y}_{jm}$ $(= \omega_{jm} \cdot \mathrm{tp}_{jm})$.

Output. A $50 + 1$ member postprocessed ensemble with hourly precipitation sums for each target location.

For stability reasons, the weights $\boldsymbol{\omega}$ are computed using values for $\hat{y}_{jm}$ and tp_{jm} rounded to two digits ($\frac{1}{100}\,\mathrm{mm\,d}^{-1}$) to avoid weights close to infinity. If $\hat{y}_{jm}$ or tp_{jm} is 0, the corresponding weight is set to 0 as well. After re-weighting, the precipitation forecasts are at the very same temporal resolution as the temperature forecasts and the rank order structure can be restored with respect to the underlying EPS (Sect. 2.3). This procedure is repeated for each target location, e.g., on a regular grid with a much finer resolution than the underlying NWP, to create high-resolution spatial predictions.

Due to the ensemble copula (Sect. 2.3) and the re-weighting procedure the full probabilistic predictions as returned by SAMOS are reduced to a $50+1$-member ensemble.

This is necessary as the precipitation postprocessing uses a censored response distribution and a parametric decomposition is not possible (central limit theorem). As a side note it has to be mentioned that the ranks of the hourly copulas are no longer strictly preserved and might sometimes differ from the original rank structure of the EPS.

2.5 New snow amount and probability of snow

Once ECC-Q and re-weighting are applied to the marginal distributions, bi-variate time series of calibrated hourly precipitation sums and 2 m temperatures are available for each of the M ensemble members. For each individual pair of member m and forecast step s the "snow indicator function" SI_{ms} can be retrieved.

$$\mathrm{SI}_{ms} = \begin{cases} \text{“dry” if} \\ \quad \text{precipitation}_{ms} \leq 0.05\,\mathrm{mm\,h^{-1}} \\ \text{“rain” if} \\ \quad \text{precipitation}_{ms} > 0.05\,\mathrm{mm\,h^{-1}} \wedge T_{2\,\mathrm{m}} > 1.2\,^{\circ}\mathrm{C} \\ \text{“snow” if} \\ \quad \text{precipitation}_{ms} > 0.05\,\mathrm{mm\,h^{-1}} \wedge T_{2\,\mathrm{m}} \leq 1.2\,^{\circ}\mathrm{C} \end{cases} \tag{11}$$

The threshold of $0.05\,\mathrm{mm\,h^{-1}}$ has been chosen as the smallest recorded value of the rain gauges used for validation is 0.1 mm. To distinguish between rain and snow we use a fixed threshold of 1.2 °C as a rough approximation, following Bellaire et al. (2011, p. 1121). The empirical probabilities π_{cs} for each of the three classes (snow, rain, and dry, which are mutually exclusive for each individual member and forecast time step) or for combinations can be computed using

$$\pi_{cs} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{1}(\mathrm{SI}_{ms} = c), \tag{12}$$

where s is a specific forecast step and c is the desired class (e.g., snow, rain, rain $\vee$ snow). $\mathbf{1}(\cdot)$ is an indicator function which takes 1 if the argument in brackets is true or 0 otherwise. The conditional expectation can be derived similarly:

$$E[c] = \frac{\sum_{i=1}^{M} \text{precipitation}_{ms} \cdot \mathbf{1}(\mathrm{SI}_{ms} = c)}{\sum_{i=1}^{M} \mathbf{1}(\mathrm{SI}_{ms} = c)}. \tag{13}$$

If one is interested in the snow height of fresh snow (E[snow] in centimeters), the snow density has to be taken into account. A rule of thumb is the "1 : 10 rule" where 1 mm of liquid water equivalent, the quantity forecasted by the postprocessing, corresponds to 1 cm of fresh snow. This is equivalent to a fresh snow density of $100\,\mathrm{kg\,m^{-3}}$. In reality, fresh snow densities can vary strongly between 10 and $526\,\mathrm{kg\,m^{-3}}$ given location and prevailing conditions (e.g., Meister, 1985; Judson and Doesken, 2000; Roebber et al., 2003). As reliable fresh snow height or density observations with the desired temporal resolution are not available for this study, a detailed verification cannot be performed. For visual representation we simply assume a mean density of $100\,\mathrm{kg\,m^{-3}}$.

3 Data

This section presents the data sets used for this study. These consist of two different EPS forecast data sets (ECMWF hindcast and operational EPS) and three different sources of observation data for model training and verification.

3.1 Numerical weather prediction data: forecast data

All predictions presented in this article are based on the ECMWF EPS. The ECMWF EPS consists of 50 perturbed ensemble members and 1 control run (50 + 1) and is initialized four times a day every 6 h. For this study, the control run is treated the same way as the 50 perturbed members. We will solely focus on the 00:00 UTC forecast run of EPS model version *43r1*. This version became operational on 22 November 2016 and the output is available at an hourly temporal resolution up to +90 h ahead on a $\sim 16\,\mathrm{km} \times 16\,\mathrm{km}$ regular longitude–latitude grid. A visual representation of the grid is shown in Fig. 1.

The presented application will focus on the winter season 2016/17 (1 December 2016 through 15 April 2017) and on predictions from +6 h to +78 h in advance, spanning the first 3 days after EPS initialization (06:00 to 06:00 UTC of 3 consecutive days).

3.2 Numerical weather prediction data: training data

To train the SAMOS models we use ECMWF hindcasts, similar to the approach of Stauffer et al. (2017b). ECMWF hindcasts become available twice a week (Mondays and Thursdays), providing a 10+1 member ensemble for the same date over the previous 20 years, initialized at 00:00 UTC. For example: on Monday 2 January 2017 hindcasts for 2 January 2016, 2015, …, 1998, and 1997 become available. As for the EPS, the hindcast control run is treated as an additional member to increase the ensemble sample size. The hindcasts are available at the same spatial resolution as the EPS, but at a 6-hourly temporal resolution only. To create the training data set for the statistical postprocessing models all hindcasts are bilinearly interpolated to each of the measurement sites (see Sect. 3.3). Overall, 235 different grid points from the numerical model are involved in the interpolation for all 199 sites.

For the statistical postprocessing methods of 2 m temperature, all 6-hourly intervals from +6 to +78 h will be used. Besides the forecasted 2 m temperature the 2 m dew point temperature, 850 hPa temperature, and surface pressure forecasts are used as additional covariates (see Sect. 4). For precipitation, 24 h total precipitation sum hindcasts are used for the forecast time steps +30, +54, and +78 h.

3.3 Observational data

Two major different observation networks will be used in the following. As in Stauffer et al. (2017b), daily liquid water equivalent observations from the Tyrol network of hydrographical services (EHYD; BMLFUW, 2018) are used for the postprocessing of daily precipitation sums. In comparison to other networks in the area, the hydrographical service maintains the highest density of stations (number of stations) with very long historical records (up to 47 years of data). The observation sites are well distributed up to an altitude of about 1800 m a.m.s.l. However, observations are only made once a day (manually) at 06:00 UTC. In the following, the observations from 110 sites in and around Tyrol are used to train the precipitation SAMOS models.

The second network consists of 89 automated weather stations operated by the national weather service (TAWES network; Zentralanstalt für Meteorologie und Geodynamik). Seventy-five out of these 89 stations provide at least 6 years of data. Observations are recorded every 10 min, of which all observations at every full hour are used for training and validation of the 2 m air temperature SAMOS models.

The TAWES network also provides automated precipitation measurements at a 10 min resolution. However, the length of historical records is much shorter compared to the time series provided by EHYD data set. Furthermore, the measurement errors of the automated rain gauges are expected to be larger than the errors from the daily manual records provided by the hydrographical service, especially during winter. Thus, we decided to not use the TAWES precipitation observations for model training and for the estimates of the spatio-temporal climatologies. Nevertheless, since observations from the hydrographical service are only available up to 2012 at this time (2018), we do use TAWES precipitation observations for validation. Therefore, daily precipitation sums are generated by taking the sum over all 10 min intervals between 06:10 and 06:00 UTC of the following day (yields 144 10 min values). Periods for which more than four 10 min values are missing are eliminated.

In addition to the temperature and precipitation observations from the hydrographical service and the TAWES network, meteorological aerodrome reports (METARs) from Innsbruck Airport are used in the verification section as it is the only longer-term source of temporally high-resolution *precipitation-phase* observations available. The weather conditions from the METARs are classified as "snow" (if the report contains SN, SG, IC, PL, SNRA, or RASN), "rain" (if the message contains DZ, RA, SNRA, or RASN), and "dry" (else). Conditions with sleet (mixed rain/snow; SNRA/RASN) are attributed to both "snow" and "rain". METARs are available every 30 min, created by either a human observer or an automated procedure if the airport is closed over night. These observations have been aggregated to an hourly temporal resolution and will be used to validate the forecasted probabilities of snowfall. Overall, 3318 observations are available for the time period of interest, with 333 cases reporting rain or sleet (10 %), 246 cases snow or sleet (7.5 %), and 2786 cases dry conditions (84 %).

Figure 1 shows an overview of the area of interest. The markers show the locations of the observational sites from the two networks (TAWES, EHYD) and the location of the airport (581 m a.m.s.l.). To the right the height distribution of the stations from the two networks is shown.

4 Statistical models

This section presents the specifications of the models that will be compared and tested in Sect. 5. During the preparation of this paper, a variety of slightly different model formulations were tested and the presented models are only a subset that was selected because they performed well or showed interesting results.

All the models follow the approaches presented in Sect. 2 but differ in their input variables and whether the data are transformed to standardized anomalies. Four models will be used for 2 m temperature and three for daily precipitation sums. The training data set to estimate the regression coefficients is composed of all forecast steps provided by the ECMWF hindcasts from +6 up to +78 h on a regular 6 h interval. For precipitation, these forecasts are aggregated to 24 h sums, resulting in forecast steps +30, +54, and +78 h. The power parameter was set to $p = 1.35$, found to have the best predictive cross-validated performance in Stauffer et al. (2017b).

Table 1 shows the different model assumptions and naming. The first two models named *EMOS* correspond to Eqs. (1)–(8) operating on the physical scale (not on standardized anomalies). One crucial modification has to be made for the 2 m temperature: interactions with factors for the time of day (*hour*; 00:00/06:00/12:00/18:00 UTC) and the station (*station*) are included to capture spatial and diurnal differences, yielding separate (and independent) coefficients for each station and each time of day. For daily precipitation sums, this extension has not been made as only 06:00 UTC observations are included (no diurnal effect required) and station-wise regression models partially returned highly unstable estimates due to the low number of observations for each individual site. Please note that the *EMOS* models are *not designed for spatial or spatio-temporal predictions* even if spatial predictions would be possible in the case of precipitation. These two models serve as a reference for the performance of the SAMOS models.

All other models are spatio-temporal (in the case of 2 m temperature) and spatial (in the case of daily precipitation sums) SAMOS models operating on the standardized anomaly scale. Thus, the spatial and temporal characteristics among all stations and for all lead times are already removed from the data and do not have to be considered in the linear predictors for location $\boldsymbol{\mu}^*$ and scale $\boldsymbol{\sigma}^*$.

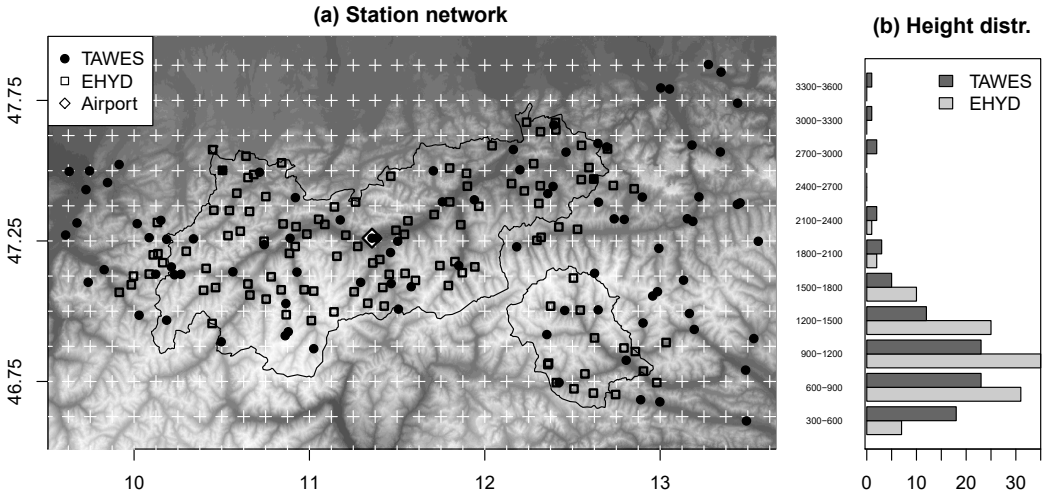

Figure 1. Panel **(a)** shows the topography of the area of interest. Overlays: center of the grid cells of the NWP model data (white crosses), governmental area of Tyrol (black outline), location of the TAWES stations (89; circles) and EHYD stations (110; squares). The airport is indicated by a diamond in the center of the map. Panel **(b)** shows the height distribution of the stations grouped into 300 m intervals: number of stations (abscissa) and altitude intervals (ordinate; meters a.m.s.l.).

The second and third pairs of models, named *SAMOS_hom* and *SAMOS_het*, are two SAMOS variations, both solely using the corresponding quantity from the EPS as a covariate (i.e., 2 m temperature and total precipitation, respectively). While *SAMOS_het* is a full heteroscedastic model including the ensemble standard deviation in the linear predictor for the scale $\boldsymbol{\sigma}^*$, *SAMOS_hom* is a homoscedastic model where the scale does not depend on any covariates. These two models allow one to quantify the improvement in the predictive performance by including the ensemble spread information in the postprocessing methods. For 2 m temperature, a fourth model called *xSAMOS_het* (*x* for *extended*) is used, which includes additional covariates for both location $\boldsymbol{\mu}^*$ and scale $\boldsymbol{\sigma}^*$. A set of multilinear models (not shown) has been tested that includes different interactions and nonlinear effects in the linear predictors, but no major improvements have been found. Thus, a relatively simple model specification for *xSAMOS_het* is included in this article to demonstrate that SAMOS can easily be extended. The multilinear *xSAMOS_het* contains three additional covariates as linear main effects for both location $\boldsymbol{\mu}^*$ and scale $\boldsymbol{\sigma}^*$. For each of the covariates separate regression coefficients are estimated during model optimization which, in this case, yields 10 coefficients in total (one intercept and four covariates in each linear predictor).

5 Results

The first two subsections show the performance of the full predictive distributions of the 2 m temperature (Sect. 5.1) and daily precipitation forecasts (Sect. 5.2). Section 5.3 shows a example of the spatial coherence restored via ECC followed by a detailed verification of hourly predictions and hourly precipitation-type classification based on the postprocessed ensembles. Last but not least, spatial forecasts for a specific forecast are shown in Sect. 5.6 to demonstrate the feasibility of high-resolution areal predictions.

5.1 Temperature (6 h intervals)

Figure 2 shows bias, continuous rank probability score (CRPS; Gneiting and Raftery, 2007), mean width of the prediction interval between the 10 % and 90 % percentiles, and CRPS skill scores, all based on the full predictive distribution returned by the statistical models. All results are temporally out-of-sample and validated on the TAWES network for all forecast steps $+6/+12/\ldots/+72/+78$ h as used to train the statistical models on hindcasts. The box-and-whiskers show station-wise mean scores for the spatio-temporal climatology (CLIM; Eq. A1), the raw EPS, and the four statistical postprocessing models (cf. Table 1).

The raw EPS performs poorly for the area of interest as the NWP model with its current spatial resolution is not able to represent the local topography. It performs even worse than the underlying climatology in terms of bias and CRPS. All statistical postprocessing models perform significantly better and are essentially bias-free. As expected, the station-wise statistical *EMOS* model performs best since it has separate model coefficients for each station location and is thus more flexible than the spatial models. In terms of CRPS, the spatial models lose about 7 %–12 % of skill (Fig. 2d; *SAMOS_hom*: −12.3 %; *SAMOS_het*: −12.3 %; *xSAMOS_het*: −6.9 %), but allow one to predict at any arbitrary location within the area of interest and any desired time between +6 and +78 h. The two models *SAMOS_hom* and *SAMOS_het* perform very

Table 1. Statistical model specification for 2 m temperature (left) and 24 h precipitation sums (right). For each model the linear predictors for μ and $\log(\sigma)$ are shown. Superscript asterisk indicate variables on the standardized anomaly scale (SAMOS). $T_{2\,\mathrm{m}}$, $\mathrm{Td}_{2\,\mathrm{m}}$, T_{850}, P, and tp are the 2 m temperature, 2 m dew point temperature, temperature in 850 hPa, surface pressure, and total precipitation ensemble forecasts respectively. $\overline{X}$ are ensemble means, $\langle X \rangle$ denote ensemble log-standard deviations. X / hour and X / station are interactions between X and the "time of the day" and/or the "station".

Models for 2 m temperature using a Gaussian response distribution.		Models for 24 h precipitation sums using a power-transformed left-censored logistic response distribution.	
Heteroscedastic EMOS models (EMOS; cf. Eqs. 1–3 and 5–7) These models *are not designed to provide spatial or spatio-temporal predictions*.			
μ	= hour / station $+\overline{T_{2\,\mathrm{m}}}$ / hour / station	μ	$= \overline{\mathrm{tp}^{1/p}} \cdot (1-z) + z$
$\log(\sigma)$	= hour / station $+\langle T_{2\,\mathrm{m}} \rangle$ / hour / station	$\log(\sigma)$	$= \langle \mathrm{tp}^{1/p} \rangle \cdot (1-z)$
Homoscedastic SAMOS models (SAMOS_hom)			
μ^*	$= \overline{T^*_{2\,\mathrm{m}}}$	μ^*	$= \overline{\mathrm{tp}^{1/p^*}} \cdot (1-z) + z$
$\log(\sigma^*)$	= constant	$\log(\sigma^*)$	= constant
Heteroscedastic SAMOS models (SAMOS_het)			
μ^*	$= \overline{T^*_{2\,\mathrm{m}}}$	μ^*	$= \overline{\mathrm{tp}^{1/p^*}} \cdot (1-z) + z$
$\log(\sigma^*)$	$= \langle T^*_{2\,\mathrm{m}} \rangle$	$\log(\sigma^*)$	$= \langle \mathrm{tp}^{1/p^*} \rangle \cdot (1-z)$
Extended Heteroscedastic SAMOS models (xSAMOS_het)			
μ^*	$= \overline{T^*_{2\,\mathrm{m}}} + \overline{\mathrm{Td}^*_{2\,\mathrm{m}}} + \overline{T^*_{850}} + \overline{P^*}$	–	
$\log(\sigma^*)$	$= \langle T^*_{2\,\mathrm{m}} \rangle + \langle \mathrm{Td}^*_{2\,\mathrm{m}} \rangle + \langle T^*_{850} \rangle + \langle P^* \rangle$	–	

similarly, indicating that the uncertainty information from the EPS 2 m temperature forecast provides barely any additional information. Small improvements can be achieved by including additional covariates (model *xSAMOS_het*).

Overall, all statistical models show promising values in terms of CRPS (median 1.45–1.65 °C) and mean absolute error (median 2.0–2.3 °C; not shown) across all four methods. The median of the mean prediction interval width for the 10 %–90 % interval is around 6.0 °C for the station-wise *EMOS* model and around 6.9–7.2 °C for the SAMOS models.

5.2 Daily precipitation sums

Figure 3 shows the verification of the daily precipitation sum predictions for the forecast steps $+30/+54/+78$ h. Again, this analysis is based on the full predictive distribution returned by the statistical models. Here, the validation is done on different stations (TAWES) than used for model fitting (EHYD; Sect. 3), so that these results are spatially and temporally out of sample. The box-and-whiskers show station-wise mean scores for the spatio-temporal climatology (CLIM; Eq. A2), the raw daily accumulated total precipitation from the ECMWF EPS (raw EPS), and the three postprocessing methods shown in Table 1.

The top row of Fig. 3 shows bias, CRPS, and the Brier score for probability of precipitation ($\mathrm{BS}_{0\,\mathrm{mm}}$). The row below shows skill scores with the raw EPS as reference. The two SAMOS models (*SAMOS_hom* and *SAMOS_het*) show the best bias among all methods but less predictive skill in terms of MAE, CRPS, and $\mathrm{BS}_{0\,\mathrm{mm}}$ than the *EMOS* model not using standardized anomalies. The distinct improvements in the $\mathrm{BS}_{0\,\mathrm{mm}}$ are expected due to the well-known wet bias of the EPS when comparing interpolated data (spatial scale) to a specific site (point scale). As for 2 m temperature, the use of the forecasted EPS uncertainty in the heteroscedastic model (*SAMOS_het*) brings barely any improvement. The performance of the *EMOS* model requires special attention. Even if this model is not designed to create spatial predictions the results show a slightly better performance than the two SAMOS models.

5.3 Spatial coherence (ensemble copula coupling)

Sections 5.1 and 5.2 examine the predictive skill of the full probabilistic predictions (SAMOS; Sect. 2.2). The next step is to apply ECC-Q based on the postprocessed hourly 2 m temperature forecasts and daily precipitation sums (Sect. 2.3) to restore the spatial structure of the forecasts.

To illustrate the effect of ECC-Q, Figs. 4 and 5 show forecasts for both 2 m temperature and daily precipitation sums, of one random member (member 38) of the forecast for 10 March 2017. Both figures show the actual forecasts of this specific member and the deviation of this member from the median of the full underlying ensemble. The latter one is used to highlight the spatial coherence which is less perceptible in the forecasts itself due to the superimpo-

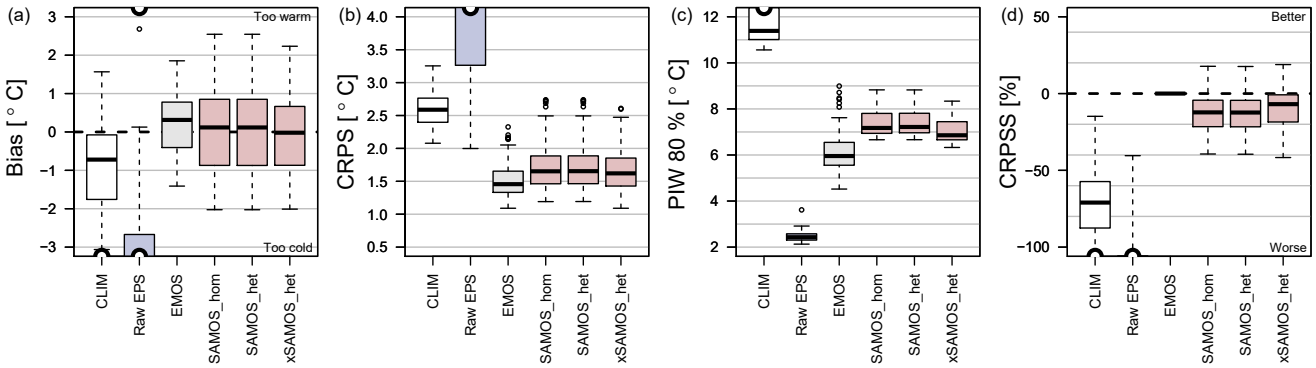

Figure 2. Scores for 2 m temperature forecasts based on the full predictive distribution based on $6/+12/\ldots/+72/+78$ h forecasts as used for model training. The box-and-whisker shows station-wise means for **(a)** bias (observation minus forecast), **(b)** CRPS, **(c)** width of the 80 % prediction interval, and **(d)** CRPS skill scores with *EMOS* as reference. Scores are shown for the climatology (CLIM), the raw EPS, and the four postprocessing models (cf. Table 1). Abscissa are set to manually specified ranges; the "semi-sphere" marker (top/bottom) indicates data outside the plotted range.

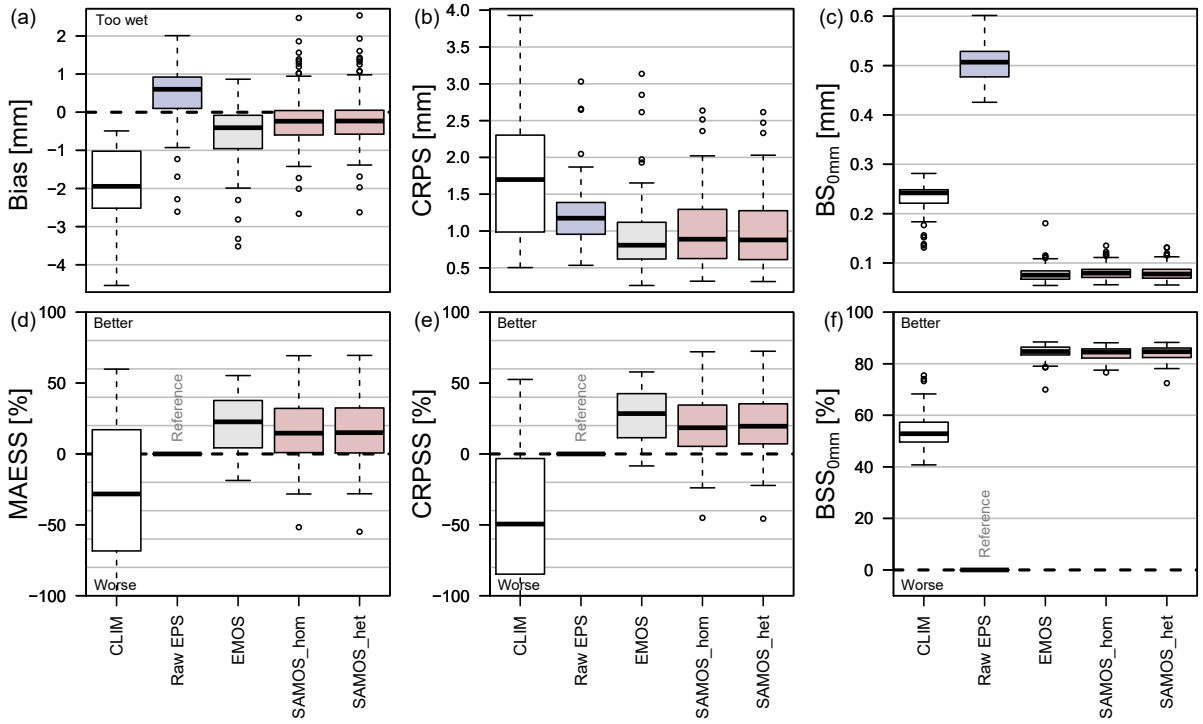

Figure 3. Scores for 24 h precipitation sums based on the full predictive distribution for +30, +54, and +78 h forecasts as used for model training. Box-and-whiskers of station-wise mean scores for **(a)** bias (observation minus forecast), **(b)** CRPS, and **(c)** Brier scores for probability of precipitation. The scores are shown for the climatology (CLIM), raw EPS, and the three postprocessing models (cf. Table 1). The lower row shows skill scores for **(d)** mean absolute error, **(e)** CRPS, and **(f)** Brier score for probability of precipitation with the *raw EPS* as reference. Positive skill scores indicate an improvement over the *raw EPS*.

sition of location- and elevation-dependent effects. Forecasts and deviations are shown for the raw ensemble, the quantiles drawn from the full probabilistic predictions, and ECC-Q after restoring the rank order structure of the EPS.

As ECC-Q uses quantiles based on equidistant probabilities, the quantiles drawn from the full probabilistic distribution are ordered. Thus, the forecasts of member 38 ($\pi = 38/52$) are always higher than the median of the ensemble

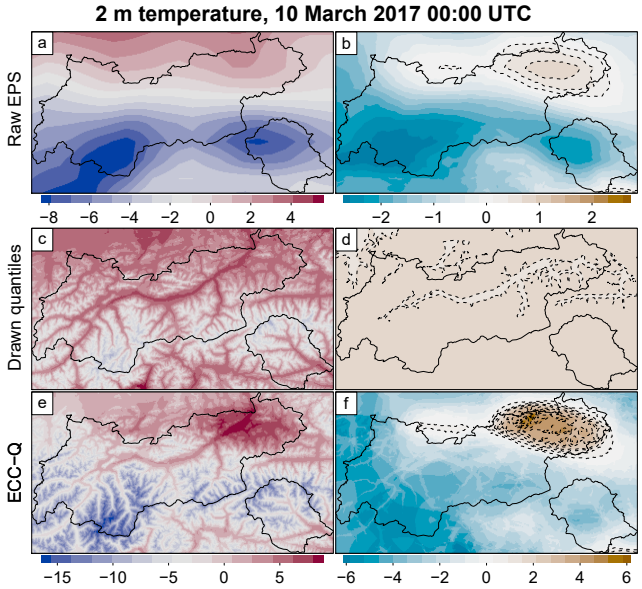

Figure 4. 2 m temperature forecasts of member 38 for 10 March 2017 00:00 UTC. Top–down: raw EPS **(a, b)**, unsorted quantile **(c, d)**, and ECC-Q **(e, f)** after restoring the rank order structure. Forecast (**a, c, e**; °C) and deviation of this forecast from the median of the corresponding ensemble (**b, d, f**; °C) are shown. Overlays: contours for positive deviation (dashed) and the borders of the governmental area of Tyrol (solid). Please note that the color scale of the top row differs from the scale of the two lower rows.

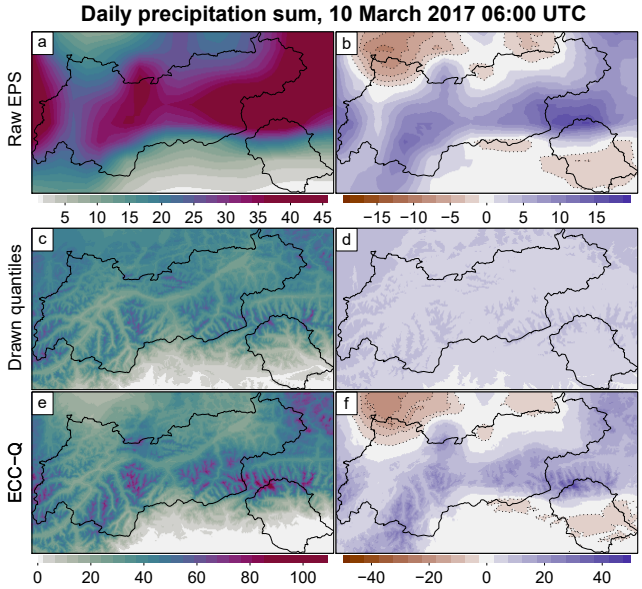

Figure 5. As Fig. 4 but for daily precipitation sums valid 9 March 2017 06:00 UTC to 10 March 2017 06:00 UTC. Forecasts **(a, c, e)** and deviations from the ensemble median **(b, d, f)** are shown in mm 24 h^{-1}. In contrast to Fig. 4 contours are plotted for negative deviations.

($\pi = 0.5$) before the rank order structure is restored. This can be seen in Figs. 4d and 5d, where the deviation against the ensemble median is plotted. In this case the deviation is (more or less) a constant positive offset across the whole domain with only little spatial structure. These small spatial features are induced by the SAMOS procedure where the data are transformed into the standardized anomaly scale and back to the physical scale (Sect. 2.2) and are not associated with the spatial coherence of the EPS (cf. Figs. 4b and 5b). To restore the spatial structure, the quantiles have to be reordered given the rank order structure of the raw EPS at each of the target locations. The bottom rows of Figs. 4 and 5 show the forecasts after rearranging the quantiles. In contrast to the non-rearranged forecasts (middle row) the postprocessed forecasts with restored rank-order structure exhibit a very similar spatial coherence to the raw EPS (top row of Figs. 4 and 5). The coherence of the EPS is maintained unchanged in large parts, but is not identical as it is slightly modified by the post-processing procedure.

5.4 Hourly temperature and precipitation sums

Sections 5.1 and 5.2 show that the postprocessing models are able to improve the predictive performance of the raw EPS for temperature and daily precipitation sums. The main goal of this study is to provide accurate and reliable snow predictions by combining hourly 2 m temperature and precipitation forecasts. Thus, an hourly temporal resolution for both temperature and precipitation forecasts is required. This section therefore shows the verification of hourly forecasts. For temperature, the hourly forecasts are based on the spatio-temporal SAMOS model *xSAMOS_het* as it shows the overall best performance among all tested spatial models. The hourly precipitation sums are based on the predictions from the *SAMOS_het* model downscaled to the desired temporal resolution using the re-weighting approach presented in Sect. 2.4. Since the re-weighted precipitation forecasts are only available as ensembles but not as full predictive distributions, ensemble verification methods are employed in the following.

Figure 6a–d show ensemble rank histograms (Hamill, 2001) for hourly temperature predictions and hourly precipitation sums for the raw EPS and the postprocessed forecasts. Each observation is assigned to a rank where observations falling below the lowest member get rank 1 and observations higher than the highest member get rank 52 (50+1 members, 52 possible ranks). A perfectly uniform distribution would indicate perfect calibration. For temperature (Fig. 6a, b), the postprocessing strongly improves calibration compared to the raw EPS. However, the pronounced U-shape indicates that the predicted uncertainty is lower than in reality (underdispersion). A similar picture can be seen for the hourly precipitation sums plotted as "stacked ensemble rank histograms" (Fig. 6c, d). The total height of the bars given the

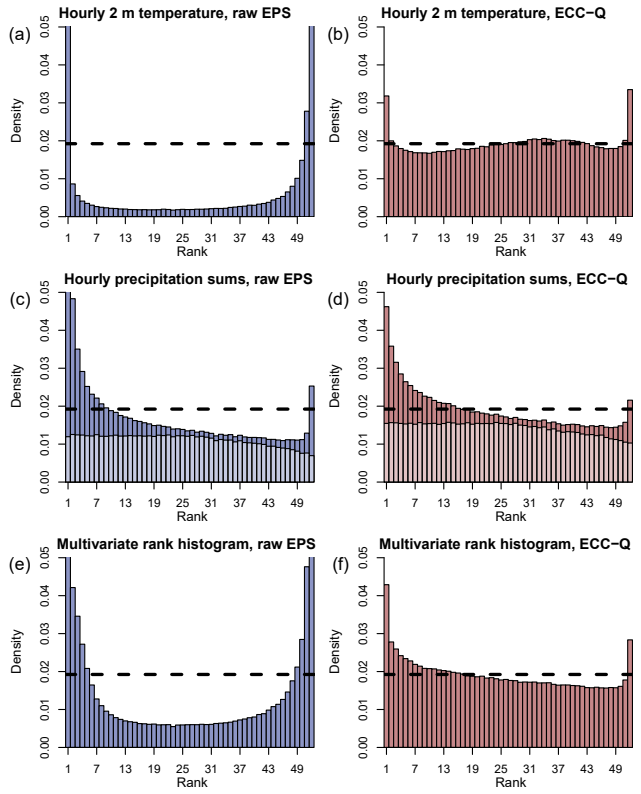

Figure 6. (Stacked) ensemble rank histograms for hourly 2 m temperature **(a, b)** and hourly precipitation sum forecasts **(c, d)** plus multivariate rank histogram **(e, f)** of the raw EPS **(a, c, e)** and postprocessed copula **(b, d, f)** with $50 + 1$ members each. The rank histograms contain all available forecasts for all stations and forecast steps $+7$, $+8$, …, $+78$ h in advance. For precipitation, the faded colors show the rank histogram for all forecasts where 50 % or more of all members predicted $0\,\mathrm{mm\,h^{-1}}$. Please note that the y axis is cut at 0.05 for all histograms.

rank shows the rank histogram of the full verification data set. The faded colors show the calibration for all forecasts where at least 50 % of all members forecasted $0\,\mathrm{mm\,h^{-1}}$ (dry cases). It can be seen that the dry cases are relatively well calibrated and that the majority of the underdispersion results from the wet cases. Nevertheless, the asymmetry (decreasing density with increasing rank) indicates a small wet bias also for the dry cases.

To score the multivariate skill of the combined temperature and precipitation forecasts, the bottom row of Fig. 6 shows multivariate (bivariate) rank histograms (Gneiting et al., 2008). In contrast to the univariate rank histograms the multivariate rank histogram takes the rank order structure between the two quantities into account. As for the univariate rank histograms the multivariate rank histogram shows much better calibration of the postprocessed predictions but shows very similar patterns to the two univariate histograms (Fig. 6a–c).

To investigate the univariate predictive performance of hourly predictions for different forecast horizons, Fig. 7 shows CRPS skill scores for all individual lead times. Each box-and-whisker contains station-wise mean skill scores over the verification period. While always on a high level, the 2 m temperature forecasts for morning hours ($+7$ to 12, $+31$ to 36, $+55$ to 60, corresponding to 07:00–12:00 UTC) show slightly less skill. For precipitation, the skill scores are overall positive but clearly decreasing with increasing forecast horizon. The lowest skill scores are found for early morning hours ($+26$ to 30, $+50$ to 54, $+74$ to 78; 02:00–06:00 UTC).

5.5 Fresh snow amounts and probability of snowfall

This section shows the verification for the main target variable. Due to the limited availability of temporally high-resolution and reliable observations this can only be done for one site, the regional airport in Innsbruck (Fig. 1). Figure 8 shows reliability diagrams (Bröcker and Smith, 2007) for the probability of precipitation (rain $\vee$ snow), rain, and snow. As Sect. 5.4 indicates that large parts of the improvements are expected to come from temperature postprocessing, three different models will be compared: the raw EPS, the full ECC-Q, and a mixed version. The mixed version uses the raw hourly precipitation forecasts from the EPS but the postprocessed temperature predictions to examine the contribution of the precipitation postprocessing. The validation for all three methods is based on the classification described in Sect. 2.5 and the aggregated METAR observations as described in Sect. 3.3.

For all three precipitation classes ECC-Q is able to outperform the raw EPS (less off-diagonal) and shows lower Brier scores and lower numbers for reliability while losing some resolution. ECC-Q is also beneficial over the mixed version using uncorrected precipitation sums. For snow the two methods using postprocessed temperature forecasts (mixed and ECC-Q) perform very similarly but show different biases. While the mixed model exhibits a wet bias (observed frequencies larger than forecasted probabilities), ECC-Q shows a dry bias. The results for snow should not be overinterpreted as snowfall is relatively rare at this station (7.5 % of all cases). The raw EPS again shows the well-known wet bias in all three classes.

Next, Fig. 9 shows a forecast time series example for a random station and a day when the temperature is just around 1.2 °C, the threshold used to decide whether the forecasted precipitation will fall as snow or rain. As no fresh snow measurements are available, a validation of the forecasted fresh snow amounts cannot be performed for this case.

What can be seen is that the ECC-Q temperature predictions (Fig. 9a) show a much larger spread than the raw EPS. The postprocessed temperature uncertainty dominates the variation of the observed temperature over the whole forecast period (days 1–3). The observations, however, nicely fall into this interval, which yields the overall well-calibrated

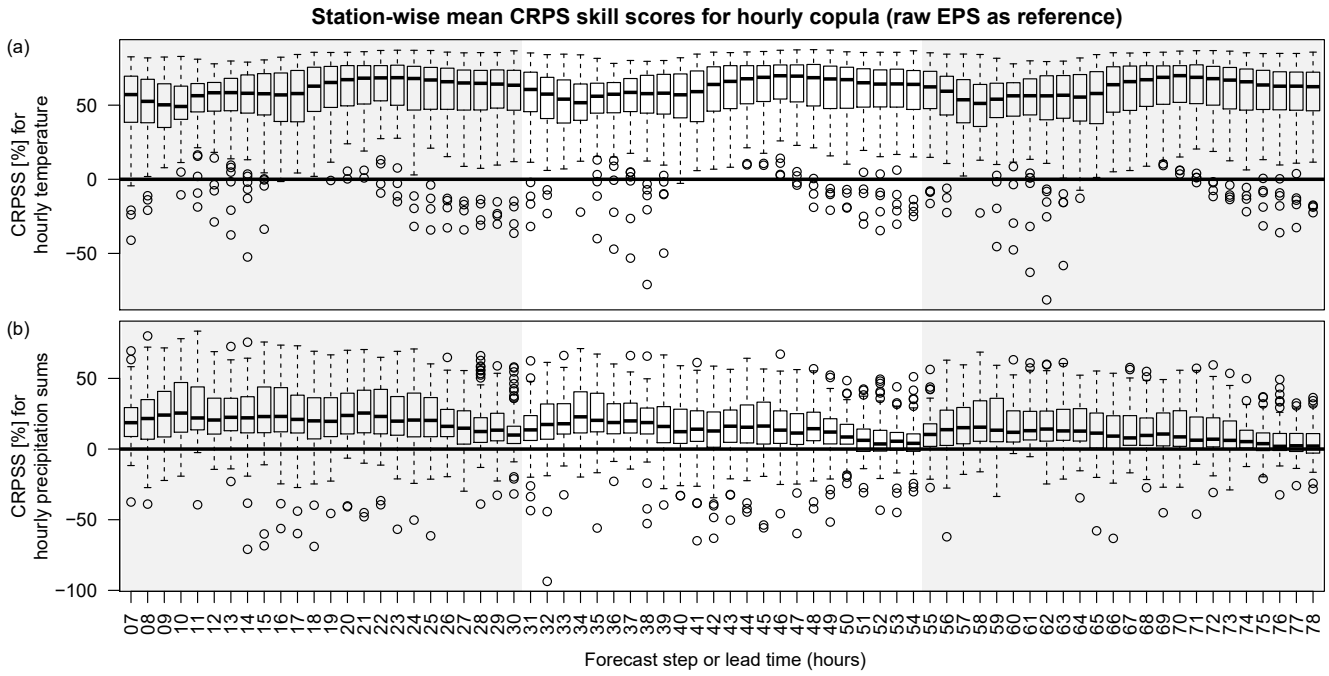

Figure 7. Continuous ranked probability skill scores (CRPSS) for 2 m temperature **(a)** and hourly precipitation sums **(b)** based on station-wise mean empirical CRPS values (50 + 1 members). The raw EPS is used as a reference. CRPSSs are shown for each individual forecast step from +7 to +78 after model initialization. CRPSSs above 0 (bold black line) show that the postprocessed hourly forecasts outperform the raw EPS.

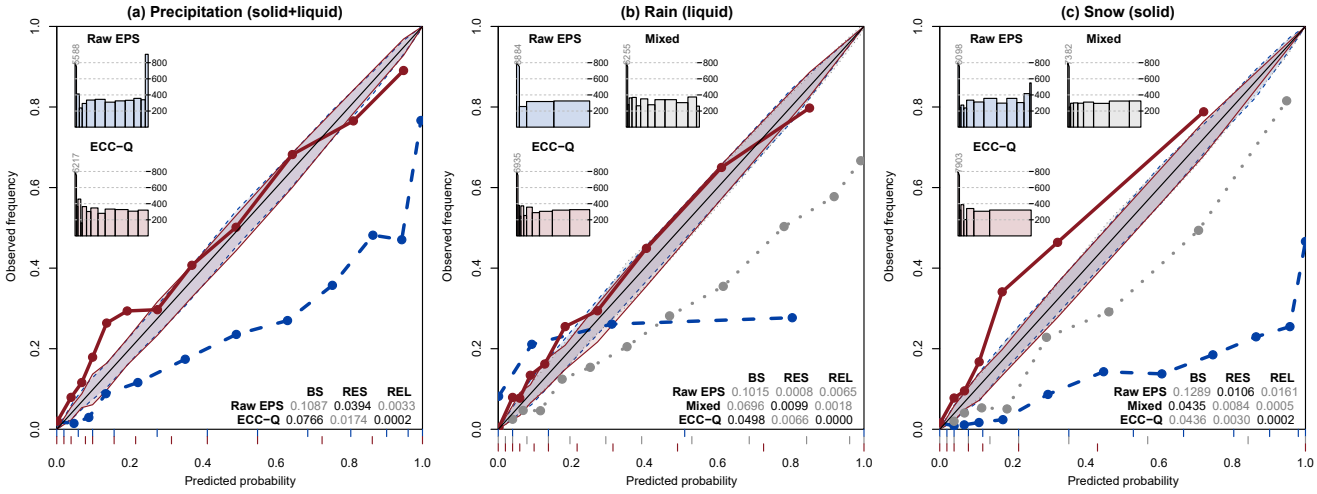

Figure 8. Reliability diagrams for hourly predictions of precipitation (snow ∨ rain; **a**), snowfall **(b)** and rain **(c)** at Innsbruck Airport based on meteorological aerodrome reports (METARs) for the raw EPS (dashed) and the postprocessed forecasts (solid). Binning based on empirical quantiles to ensure a similar number of observations per bin (bins indicated along the x axis). The shaded area shows the 90 % confidence interval. Histograms: counts of the number of observations in each bin in the reliability diagram. The analysis is based on ≈ 9700 observation–forecast pairs for each precipitation type. Mean Brier score (BS), as well as mean resolution (RES) and reliability (REL) from a BS decomposition (Murphy, 1973), are shown in the lower right corner.

forecasts (see Fig. 6). For precipitation (Fig. 9c), the differences between the raw EPS and the postprocessed copula are less pronounced. Fig. 9b shows the probability of snow ∨ rain (precipitation), rain, and snow as defined by Eq. (12). The expected amounts of snow ∨ rain (precipitation) and snow from the postprocessed forecasts are plotted in Fig. 9d. Rather than plotting each individual ECC member, the median and two confidence intervals are shown. For this specific date and

location, the median shows 30.5 mm of precipitation (rain and/or snow liquid water equivalent) accumulated over the 3 consecutive days, of which 8.4 mm is expected to fall as snow. When assuming the 1 : 10 rule (Sect. 2.5) and not taking the alteration of the aging snow into account, this corresponds to a median of 8.4 cm of fresh snow within 3 days.

5.6 Spatial forecast example

As a last result, Figs. 10 and 11 show a spatial forecast example to demonstrate the ability to create high-resolution spatial predictions. These results show the +48 h forecast initialized 8 March 2017 on an approximately 500 m × 500 m grid (corresponds to the +48 h forecast shown in Fig. 9).

While Fig. 10 shows the probability of precipitation (snow ∨ rain), rain, and snow, Fig. 11 shows the expected amount of precipitation for the period > 47 to +48 h. The color coding represents the dominant precipitation type based on $\pi_{\text{snow, +48 h}}$ and $\pi_{\text{rain, +48 h}}$ (cf. Eq. 12). In addition, the snow line ($\pi_{\text{snow, +48 h}} > \pi_{\text{rain, +48 h}}$) is shown. For visual purposes the spatial predictions are plotted for the whole domain even if parts of the area are already outside the area covered by the stations used to create the underlying observation climatologies and to train the statistical models. Thus, forecasts outside the dashed line (Fig. 10a) should be interpreted with caution. The individual EPS and ECC-Q members used to derive probabilities and the expectation can be found in Appendix B; one specific member is shown in more detail in Sect. 5.3.

6 Discussion

This article presents a new hybrid approach to combine standardized anomaly output statistics (SAMOS) with ensemble copula coupling (ECC) and a novel re-weighting scheme for probabilistic snow forecasts. The results demonstrate that the new approach provides a framework for accurate high-resolution spatio-temporal probabilistic forecasts for 2 m temperature, precipitation, and snowfall over complex terrain.

The use of ECMWF hindcasts for model training and ECMWF EPS for prediction offers a computationally efficient way to get the required inputs for the SAMOS method (see Appendix A). Rather than estimating a complex spatio-temporal climatology for each covariate (as in Dabernig et al., 2017), only empirical moments (mean and standard deviation) of an appropriate hindcast subset have to be derived. The latest eight hindcast runs (4 weeks) centered around the date of interest are used to capture the seasonality. As this processing step is very cheap in terms of computational costs, one can easily derive hindcast climatologies for a range of possible covariates, which allows for a simple and low-cost multilinear extension of the SAMOS approach. Furthermore, due to the use of a rolling 4-week training period, the post-processing procedure automatically adapts itself to possible changes in the underlying NWP model within a few weeks. However, the rank histograms (Fig. 6) for both the 2 m temperature and daily precipitation sums show a pronounced U-shape. The same characteristics can be seen for all tested postprocessing models (not shown) whether or not standardized anomalies are used. The rank histograms for in-sample predictions based on the training data set itself (not shown) do not show this distinct pattern. A possible reason could be that the forecasted uncertainty of the hindcasts and the uncertainty information from the current EPS seem to differ. If the EPS overall provided sharper forecasts than the hindcast on which the regression coefficients are estimated, this would also yield underdispersive predictions after postprocessing. A detailed analysis of this specific issue was performed (beyond this article; not shown), but a clear statement to prove or falsify the hypothesis cannot be given.

The additional ensemble copula coupling (ECC-Q; Sect. 2.3) and re-weighting strategy yield satisfying results and are able to restore the spatial coherence based on the spatial structure of the raw EPS (Sect. 5.3, Appendix B). However, the bivariate verification (Fig. 6) shows distinct underdispersion. Additional tests have been performed to verify the improvement by restoring the multivariate rank order structure. Therefore, the multivariate rank histogram has been computed using random correlation by drawing a random rank order structure from the ensemble. It turns out (not shown) that the multivariate rank histogram with the random rank order structure only differs marginally from the one shown in Fig. 6 for both the raw EPS and ECC-Q. In other words: the correlation between 2 m temperature and hourly precipitation sums is negligible, at least for this study. Thus, the impacts of the cases where the rank order structure is not strictly preserved due to the re-weighting (Sect. 2.4) are not further investigated as no verifiable effect is expected.

Nevertheless, the method is still able to strongly improve calibration and reliability of the forecasts, especially for 2 m temperature, even though the sharpness is rather low. The mean 80 % prediction interval width for temperature is between 6.9 and 7.2 °C for the SAMOS methods. On a rainy/snowy day this interval is quite likely wider than the overall diurnal temperature variation. The relatively wide predictive intervals are a result of the input data. Due to the current spatial resolution, the EPS is not able to represent the area of interest in all its details. Consequently, a wide range of local features are not yet included. To mention one specific feature: the EPS shows a far-too-strong near-surface cooling over night, especially over snow. Errors of 15 °C between the forecasted 2 m temperature and the corresponding observation are relatively frequent for Alpine grid points. Furthermore, the forecasted EPS uncertainty does not seem to be very informative as almost no improvements can be seen when including it in the statistical models.

To improve the temperature forecasts, we include the temperature from the 850 hPa level as an additional covariate, which can be seen as a "free atmosphere" prediction over the

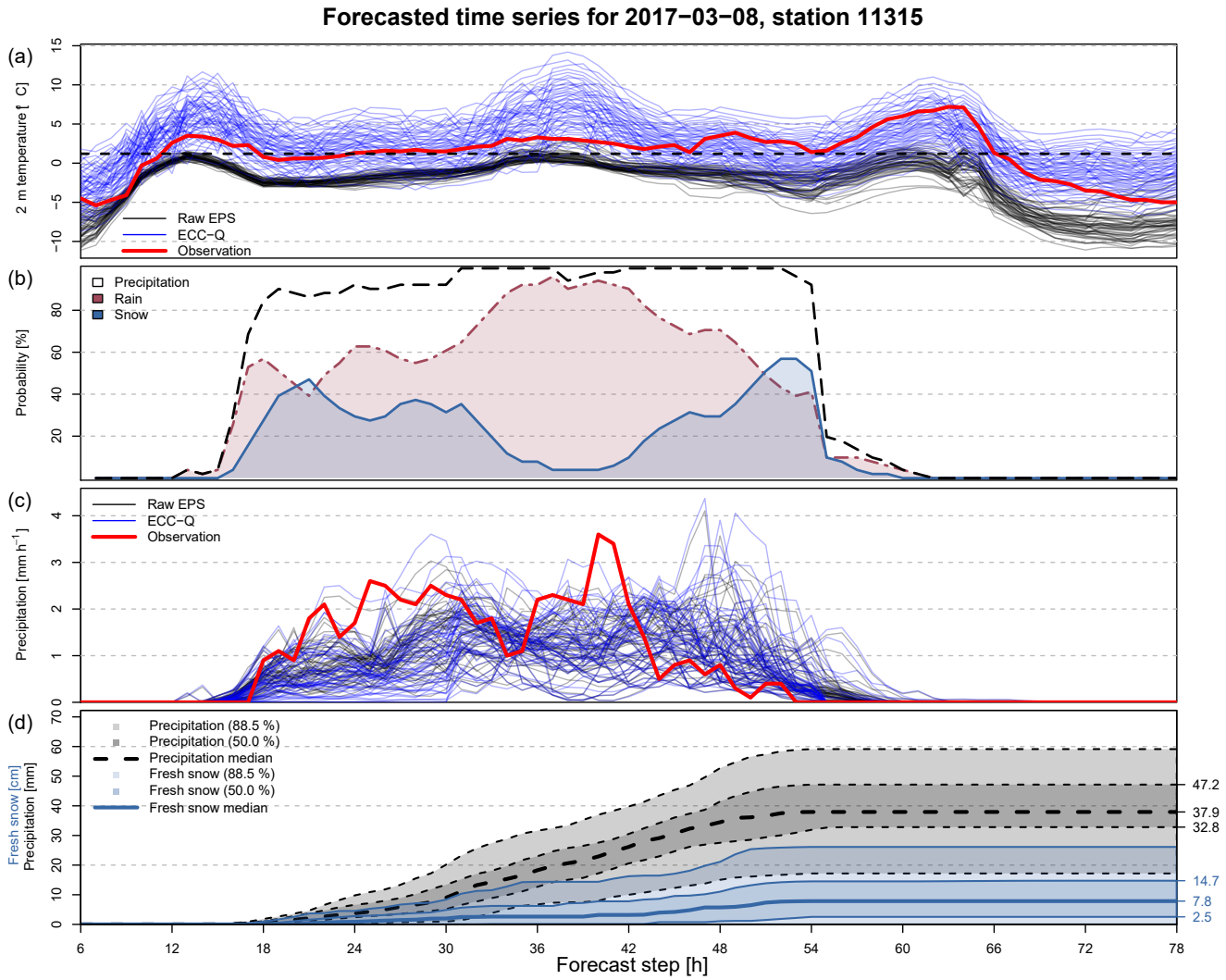

Figure 9. Example prediction for 8 March 2017 (station 11315, Holzgau) for the whole forecast horizon +6 up to +78 h ahead. **(a)** Raw EPS forecast (black), postprocessed copula (blue), and observation (red; bold) for 2 m temperature. The black dashed line is the 1.2 °C line used for precipitation type classification. **(b)** Probability of snow (blue solid), rain (red dotdash), and precipitation (snow ∨ rain; black dashed). **(c)** Hourly precipitation forecasts and observations as in panel **(a)**. **(d)** Postprocessed forecasts for precipitation sum (dashed; mm), and fresh snow height (solid; cm) using the 1 : 10 rule (snow density of 100 kg m^{-3}). Predicted medians, predicted 50 % intervals, and predicted 88.5 % intervals are shown.

area of interest. Furthermore, the 850 hPa temperature is a prognostic quantity which should be less strongly affected by possibly unrealistic surface processes (cooling/heating effects). In addition, surface pressure and 2 m dew point temperature are included to correct for weather-situation-dependent errors and very dry/wet conditions. The model shown in this article only includes the additional covariates as linear main effects and is more a proof of concept. We have also tested derived covariates such as 2 m potential temperature and nonlinear mixtures of 2 m temperature and 850 hPa temperatures to allow high-elevation stations to take the information from an elevated air mass ("free atmosphere") rather than from the near surface. As none of these models showed large improvements, and for simplicity, we decided not to show the results of these more complex models in this article. However, the "extended heteroscedastic SAMOS model" demonstrates that the SAMOS model can easily be extended by including additional covariates which do not necessarily have to be linear. As shown, this allows one to further improve the predictive performance, even with this simple model. A more flexible SAMOS model might bring further improvements, e.g., by including a larger set of covariates, including interactions between the different covariates, or by using more flexible effects such as multi-dimensional effects which can be used to represent elevation-

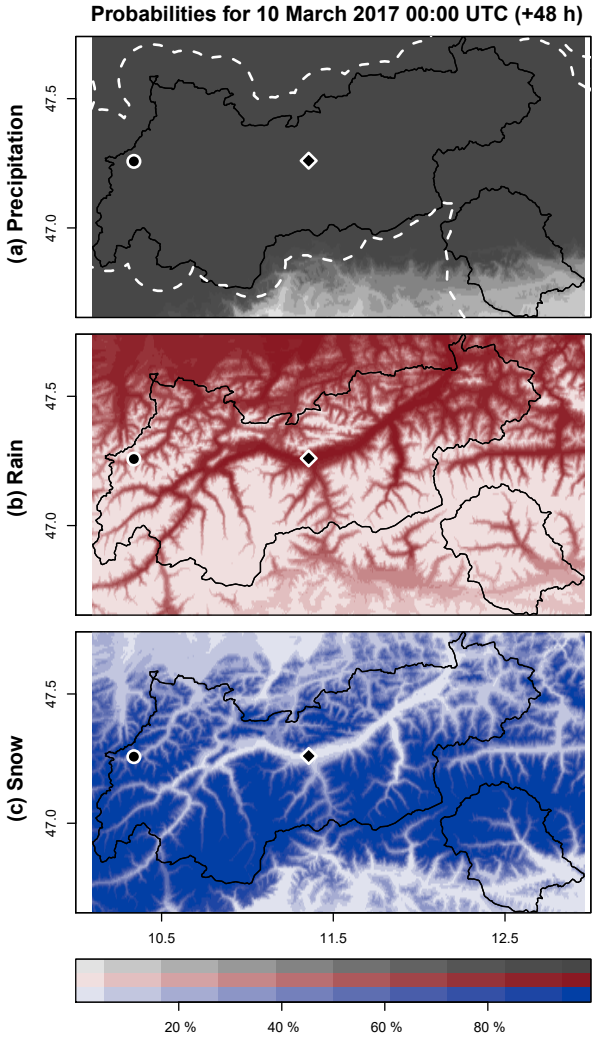

Figure 10. Top–down: 1 h probability of precipitation (rain ∨ snow), rain, and snow for 10 March 2017 00:00 UTC (+48 h forecast initialized 8 March 2017). Overlays: the governmental area of Tyrol (solid line), Innsbruck Airport (diamond), and the location of the example station used in Fig. 9 (circle). The white dashed line outlines the area not further away than 10 km from the closest measurement site.

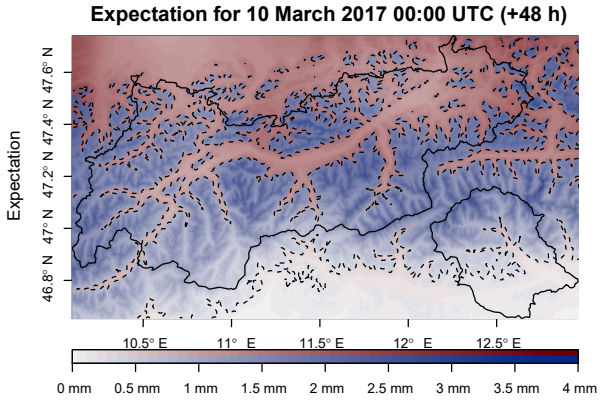

Figure 11. Expected 1 h amount of liquid water content for 10 March 2017 00:00 UTC (+48 h forecast initialized 8 March 2017). Areas with a higher chance of observing snow are shown in blue, those with a higher chance of observing rain in red. The dashed line (top) shows the forecasted snow line with an equal chance of observing snow or rain (Eq. 12). Overlay: governmental area of Tyrol (solid line).

dependent effects and which will be worth investigating in more detail in the future.

As the results show (Fig. 3), the *EMOS* model for daily precipitation sums slightly outperforms the SAMOS models, which is somehow unpleasant. A possible reason is that the overall (not location-dependent) bias and slope correction is of most importance and that this simple model is better able to correct for it. A second reason could be that the underlying observation climatology (which is an all-year climatology; Appendix A) might not perfectly capture the cold season and causes the slightly worse predictive performance of the SAMOS models. Further improvements of the underlying climatology might be beneficial for the predictive skill of the SAMOS results.

One of the biggest advantages of the proposed hybrid approach is that forecasts can be produced on the same temporal scale as the current EPS even if the underlying data sets used for model training (hindcasts and observations) are available on coarser temporal scales or even different timescales for different variables. This allows one to combine the best information from (location-)independent sources to get the most reliable probabilistic predictions possible. For the present study, two observation networks have been combined, one providing long-term daily precipitation records, and one providing temporally highly resolved temperature measurements.

Overall the 2 m temperature and precipitation forecasts serve as a good proxy for probabilistic snowfall forecasts, which is the main target variable of this study. The results show very promising results in terms of calibration and reliability of both the expected amount of precipitation and fresh snow, but also the probability of observing snowfall at an hourly temporal resolution.

Code and data availability. The main parts of this study are based on *R* package bamlss (Umlauf et al., 2017) to compute the spatio-temporal observation climatologies and *R* package crch (Messner et al., 2016) to estimate the (censored) non-homogeneous regression models. The continuous ranked probability scores are based on *R* package scoringRules (Jordan et al., 2018).

Observations from the hydrographical service (BMLFUW, 2018) can be downloaded from the website of the Bundesministerium für Land und Forstwirtschaft und Wasserwirtschaft (http://ehyd.gv.at, last access: 14 November 2018).

Appendix A: Standardized anomaly model output statistics (SAMOS)

For spatio-temporal ensemble postprocessing we followed the approach of Dabernig et al. (2017) and Stauffer et al. (2017b), which we summarize in the following. In contrast to other statistical postprocessing methods, SAMOS uses standardized anomalies for both the response and the covariates. This allows one to remove location-specific and time-specific characteristics from the data and to estimate one single regression model for all stations and forecast lead times at once. For this study we closely follow the original articles (Dabernig et al., 2017; Stauffer et al., 2017b) but slightly modify the specification, especially for the temperature SAMOS, to adapt to the different study area.

Observation climatologies. Two separate spatio-temporal models have been estimated for 2 m air temperature observations and daily precipitation sums. Both models have effects to capture seasonal, altitudinal, and spatial climatological features represented by (multi-dimensional) nonlinear functions. The 2 m temperature observations are available at an hourly temporal resolution. Therefore, additional nonlinear cyclic effects have to be included to capture the diurnal effects in the climatological estimates.

The spatio-temporal model for the 2 m temperature uses the geographical location (longitude lon, latitude lat, and altitude alt), the "hour of the day" (hour), and the "day of the year" (doy) as covariates and is specified as follows:

$$
\begin{aligned}
&\text{temperature} \sim \mathcal{N}\left(\widetilde{\boldsymbol{\mu}}_y, \widetilde{\boldsymbol{\sigma}}_y\right),\\
&\widetilde{\boldsymbol{\mu}}_y = f_1(\text{hour}, \text{doy}, \text{alt})\\
&\quad + f_2(\text{hour}, \text{doy}) + f_3(\text{doy}, \text{lon}, \text{lat})\\
&\quad + f_4(\text{hour}) + f_5(\text{doy}) + f_6(\text{doy}, \text{alt}) + f_7(\text{alt}),\\
&\log\left(\widetilde{\boldsymbol{\sigma}}_y\right) = g_1(\text{hour}, \text{doy}, \text{alt}) + g_2(\text{hour}, \text{doy})\\
&\quad + g_3(\text{doy}, \text{lon}, \text{lat})\\
&\quad + g_4(\text{hour}) + g_5(\text{doy}) + g_6(\text{doy}, \text{alt}) + g_7(\text{alt}),
\end{aligned}
\tag{A1}
$$

where $f_\bullet$ and $g_\bullet$ are up to three-dimensional smooth spline effects. Cyclic P-splines are used for all effects depending on the "day of the year" or the "hour of the day"; all other effects use penalized thin plate splines with a varying number of possible degrees of freedom. Following the same concept, the spatio-temporal model for daily precipitation sums is defined as

$$
\begin{aligned}
&\text{precipitation}^{1/p} \sim \mathcal{L}_0\left(\widetilde{\boldsymbol{\mu}}_y, \widetilde{\boldsymbol{\sigma}}_y\right),\\
&\widetilde{\boldsymbol{\mu}}_y = f_1(\text{alt}) + f_2(\text{doy}) + f_3\left(\text{lon}, \text{lat}\right)\\
&\quad + f_4(\text{doy}, \text{lon}, \text{lat}),\\
&\log\left(\widetilde{\boldsymbol{\sigma}}_y\right) = g_1(\text{alt}) + g_2(\text{doy}) + g_3(\text{lon}, \text{lat})\\
&\quad + g_4(\text{doy}, \text{lon}, \text{lat}).
\end{aligned}
\tag{A2}
$$

As for Eq. (A1), cyclic P splines are used for effects which depend on the "day of the year", while all others use penalized thin plate splines. The major difference to the temperature climatology (Eq. A1) is that a left-censored logistic response distribution $\mathcal{L}_0$ is used on power-transformed observations of precipitation$^{1/p}$ ($p = 1.35$; cf. Stauffer et al., 2017b). The complexity of the linear predictors in Eq. (A2) is lower than in Eq. (A1) as no effects for diurnal variation have to be considered.

Model climatology. Similar spatio-temporal climatologies as for the observations could be estimated for all quantities from the EPS which are used as covariates in the SAMOS models. This would have to be done for each quantity separately using a reasonably large data set of historical EPS forecasts. However, we instead extract the model climatologies directly from ECMWF hindcasts. These hindcasts are produced operationally twice a week and consist of $10 + 1$ members using the same model version and model specification as the current EPS. For each hindcast run the forecasts for the same date over the most recent 20 years are computed. The hindcasts are designed to represent the climatology of the current EPS model and are used to calibrate EPS forecasts and as input for postprocessing applications (e.g., Hagedorn et al., 2012, 2008). For our SAMOS approach we can thus simply derive the empirical mean and empirical standard deviation over a set of hindcasts to get the climatological estimates $\widetilde{\boldsymbol{\mu}}_x$ and $\widetilde{\boldsymbol{\sigma}}_x$ required to compute the standardized anomalies for covariate $\mathbf{x}$ (Eq. 9). Climatologies for lead times when no hindcast output is available (between the regular 6 h interval) are created using simple grid-point-wise linear interpolation.

Hindcasts are produced every Monday and Thursday (available Tuesday/Friday), computed 2 weeks in advance. Taking hindcasts for ± 2 weeks around the date of interest yields eight independent hindcast runs with 11 members and 20 years of (re-)forecasts each, which yields $8 \cdot 11 \cdot 20 = 1760$ forecasts. With this large number of independent predictions these climatological estimates are fairly robust. Due to the 4-week centered rolling window the climatologies automatically adapt themselves to the prevailing season. Separate climatologies for each forecast step are required to capture diurnal cycles (for temperature) and to account for changes in the model climate with increasing forecast horizon such as drifting means or increasing ensemble standard deviation. Thus, for this study, 13 separate climatologies for the temperature models ($[+6\,\text{h}, +12\,\text{h}, \ldots, +72\,\text{h}, +78\,\text{h}]$) and 3 climatologies for the precipitation forecasts ($[+30\,\text{h}, +54\,\text{h}, +78\,\text{h}]$) are required.

Estimation of the SAMOS models (see Table 1). Equations (1)–(3) and (5)–(7) show the basic heteroscedastic models used for *SAMOS_hom* and *SAMOS_het*. The only modification is to replace the response $\boldsymbol{y}$ and the covariate $\mathbf{x}$ with the corresponding standardized anomaly $\boldsymbol{y}^*$ and $\mathbf{x}^*$ (Eq. 9). For

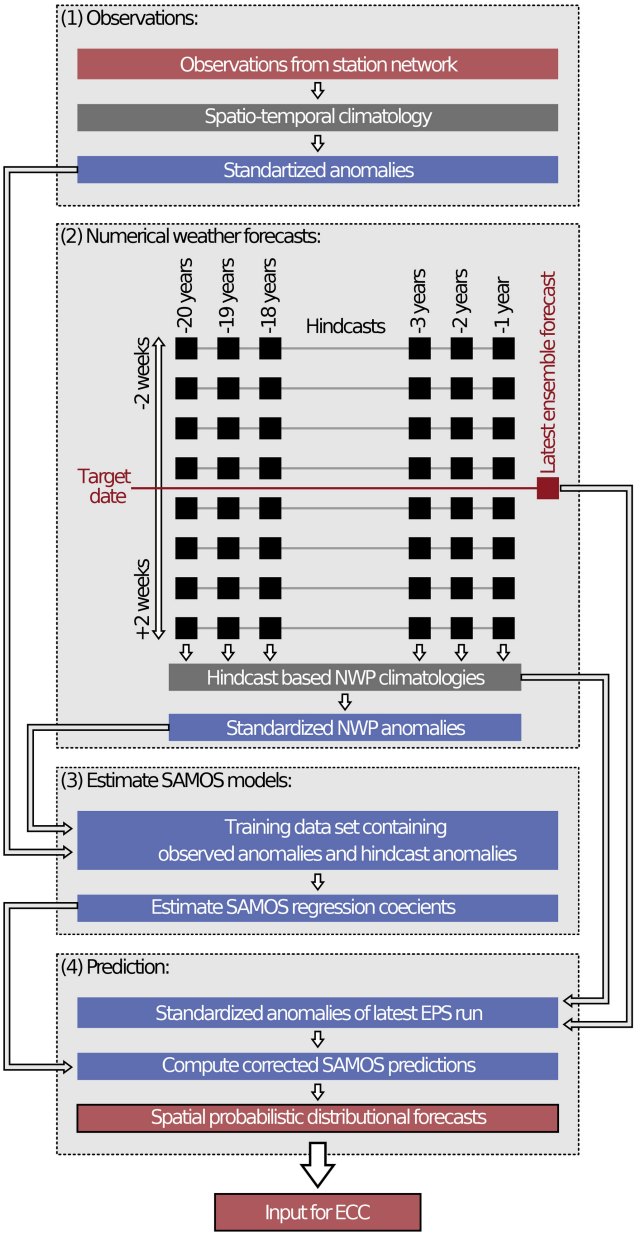

Figure A1. Schematic concept of the SAMOS postprocessing based on ECMWF hindcasts (black), ECMWF EPS forecasts (red), and observations (orange). Background climatologies (gray) are used to convert the data from the physical scale into standardized anomalies (blue) used to estimate the regression coefficients of the SAMOS postprocessing method. The SAMOS correction can be applied to the standardized anomalies of a new EPS forecast to obtain spatial or spatio-temporal probabilistic forecasts (full distribution). These results are used as input for the ECC approach.

the *xSAMOS_het* model the linear predictors in Eqs. (1)–(3) are extended by simply adding additional covariates, resulting in a multilinear SAMOS model.

Once the regression coefficients of the SAMOS model have been estimated, future ensemble forecasts can be corrected by first computing standardized anomalies using the same model climatology as for model training and correcting the standardized anomalies of the ensemble forecast using the estimated SAMOS models. As the outcomes (μ_i^* and σ_i^*) are on the standardized anomaly scale, they have to be rescaled with respect to the observation climatology to obtain physical values (e.g., °C or mm). The final predictive distribution is thus

$$y_i \sim \mathcal{D}\left(\mu_i^* \cdot \widetilde{\sigma}_{y,i} + \widetilde{\mu}_{y,i},\ \sigma_i^* \cdot \widetilde{\sigma}_{y,i}\right)^p, \tag{A3}$$

where $\mathcal{D}$ represents the normal distribution $\mathcal{N}$ in the case of 2 m temperature postprocessing with $p = 1$ and $\mathcal{L}_0$ in the case of the power-transformed daily precipitation sums' postprocessing with $p = 1.35$.

Algorithm 1 presents pseudo-code for all steps. The same is shown in Fig. A1 as a graphical representation of this procedure, visualizing the required data sets, the required steps, and their dependencies.

Appendix B: Individual copula members

Figures B1 and B2 show the individual EPS members (Fig. B1) and the corresponding re-weighted ensemble copula coupling members (Fig. B2) for the +48 h forecast for 10 March 2017 10:00 UTC as used to derive the probabilities and expectation plotted in Figs. 10 and 11. For easier comparison the NWP forecasts are bilinearly interpolated to $\sim 500 \times 500\ \mathrm{m}^2$ to match the resolution of the postprocessed predictions.

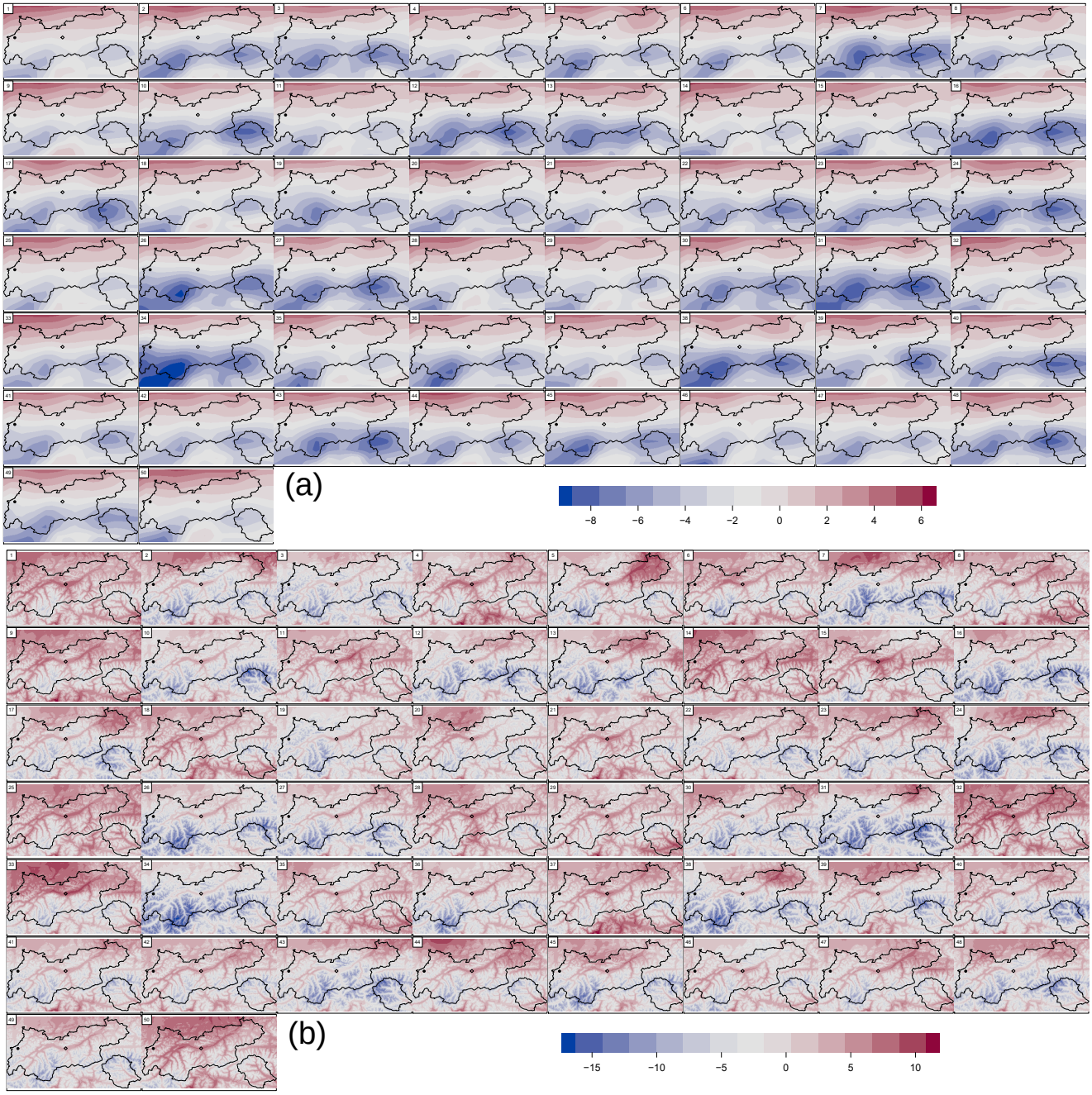

Figure B1. Stamps for +48 h forecasts initialized 8 March 2017 00:00 UTC (valid for 10 March 2017 00:00 UTC). Individual EPS members for 2 m temperature **(a)** and the corresponding copula members **(b)**. Please note that the color scale for all members of one type (EPS/copula) is identical, but the scales between the raw EPS and the results from the postprocessing differ.

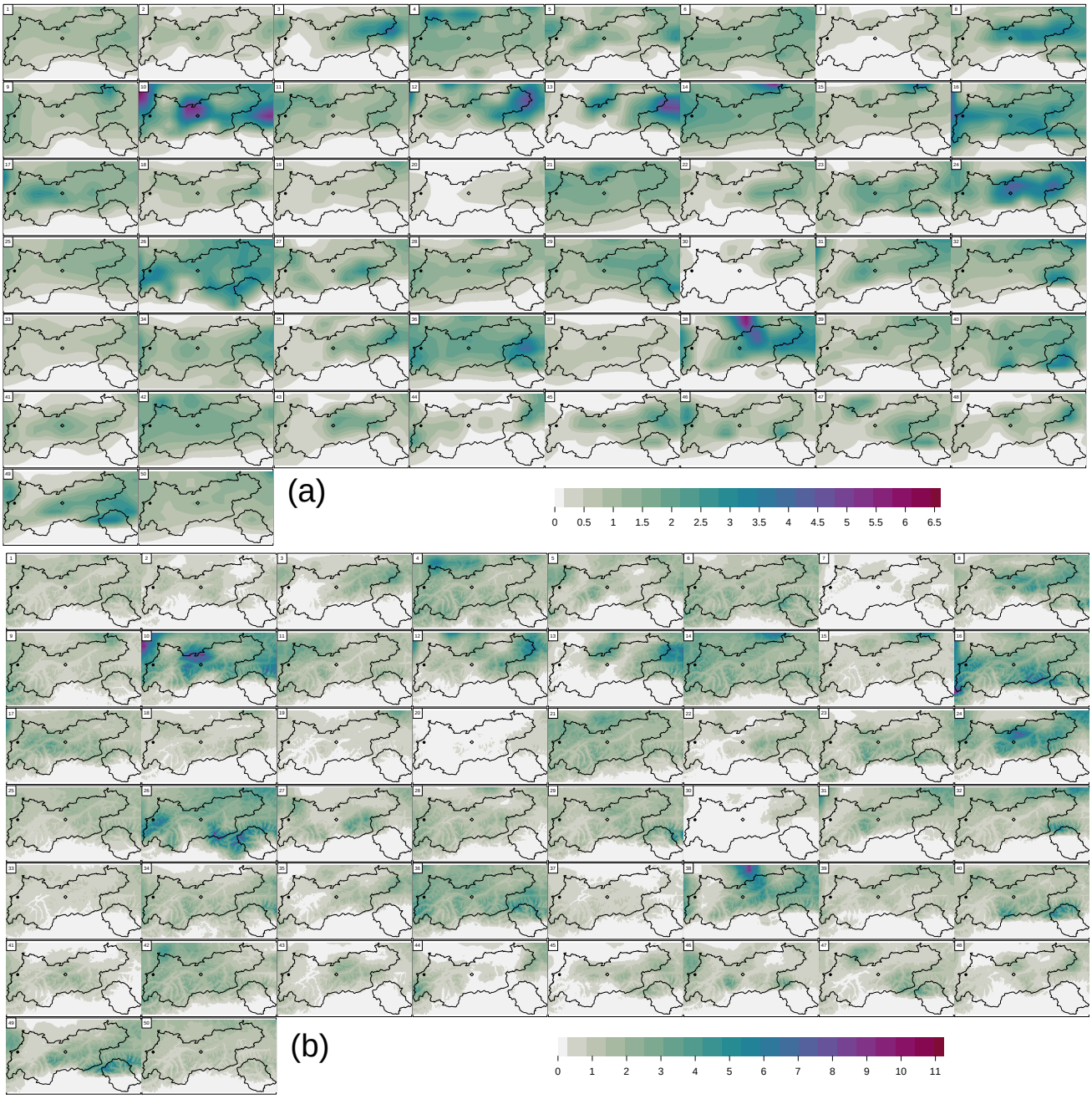

Figure B2. Stamps for +48 h forecasts initialized 8 March 2017 00:00 UTC (valid for 10 March 2017 00:00 UTC). Individual EPS members for 1 h precipitation sums **(a)** and the corresponding copula members **(b)**. Please note that the color scale for all members of one type (EPS/copula) is identical, but the scales between the raw EPS and the results from the postprocessing differ.

Author contributions. This study summarizes the ideas developed within our most recent research project by all the members, including RS, GJM, JWM, and AZ. The majority of the work for this study was performed by RS. The statistical models are, to a large extent, based on the two R packages bamlss and crch developed by JWM and AZ (and others). All the authors closely worked together discussing the results and findings and commented on this paper.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This project was partially funded by the Austrian Science Fund (FWF), grant TRP 290, and the Austrian Research Promotion Agency (FFG), grant no. 858537. The data sets are provided by the Zentralanstalt für Meteorologie und Geodynamik Vienna (ZAMG; https://zamg.ac.at, last acccess: 14 November 2018) and the Federal Ministry of Agriculture, Forestry, Environment and Water Management (BMLFUW), Abteilung IV/4 – Wasserhaushalt (http://ehyd.gv.at, last acccess: 14 November 2018).

Edited by: Christopher Paciorek
Reviewed by: two anonymous referees

References

Amt der Tiroler Landesregierung: Statistisches Jahrbuch Bundesland Tirol, https://www.tirol.gv.at/fileadmin/themen/statistik-budget/statistik/downloads/Statistisches_Handbuch_2014.pdf (last access: 15 July 2016), 2014.

Bellaire, S., Jamieson, J. B., and Fierz, C.: Forcing the snow-cover model SNOWPACK with forecasted weather data, The Cryosphere, 5, 1115–1125, https://doi.org/10.5194/tc-5-1115-2011, 2011.

BMLFUW: Bundesministerium für Land und Forstwirtschaft, Umwelt und Wasserwirtschaft (BMLFUW), Abteilung IV/4 – Wasserhaushalt, available at: http://ehyd.gv.at, last access: 14 November 2018.

Bouallègue, Z. B. and Theis, S. E.: Spatial Techniques Applied to Precipitation Ensemble Forecasts: from Verification Results to Probabilistic Products, Meteorol. Appl., 21, 922–929, https://doi.org/10.1002/met.1435, 2014.

Bröcker, J. and Smith, L. A.: Increasing the Reliability of Reliability Diagrams, Weather Forecast., 22, 651–661, https://doi.org/10.1175/WAF993.1, 2007.

Dabernig, M., Mayr, G. J., Messner, J. W., and Zeileis, A.: Spatial Ensemble Post-Processing with Standardized Anomalies, Q. J. Roy. Meteor. Soc., 143, 909–916, https://doi.org/10.1002/qj.2975, 2017.

Fraley, C., Raftery, A. E., and Gneiting, T.: Calibrating Multimodel Forecast Ensembles with Exchangeable and Missing Members Using Bayesian Model Averaging, Mont. Weather Rev., 138, 190–202, https://doi.org/10.1175/2009MWR3046.1, 2010.

Gebetsberger, M., Messner, J. W., Mayr, G. J., and Zeileis, A.: Fine-Tuning Non-Homogeneous Regression for Probabilistic Precipitation Forecasts: Unanimous Predictions, Heavy Tails, and Link Functions, Mon. Weather Rev., 145, 4693–4708, https://doi.org/10.1175/MWR-D-16-0388.1, 2017.

Gneiting, T. and Raftery, A. E.: Strictly Proper Scoring Rules, Prediction, and Estimation, J. Am. Stat. Assoc., 102, 359–378, https://doi.org/10.1198/016214506000001437, 2007.

Gneiting, T., Raftery, A. E., Westveld III, A. H., and Goldman, T.: Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation, Mon. Weather Rev., 133, 1098–1118, https://doi.org/10.1175/MWR2904.1, 2005.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L., and Johnson, N. A.: Assessing Probabilistic Forecasts of Multivariate Quantities, with an Application to Ensemble Predictions of Surface Winds, Test, 17, 211–235, https://doi.org/10.1007/s11749-008-0114-x, 2008.

Hagedorn, R., Hamill, T. M., and Whitaker, J. S.: Probabilistic Forecast Calibration Using ECMWF and GFS Ensemble Reforecasts. Part I: Two-Meter Temperatures, Mon. Weather Rev., 136, 2608–2619, https://doi.org/10.1175/2007MWR2410.1, 2008.

Hagedorn, R., Buizza, R., Hamill, T. M., Leutbecher, M., and Palmer, T. N.: Comparing TIGGE Multimodel Forecasts with Reforecast-Calibrated ECMWF Ensemble Forecasts, Q. J. Roy. Meteor. Soc., 138, 1814–1827, https://doi.org/10.1002/qj.1895, 2012.

Hamill, T. M.: Interpretation of Rank Histograms for Verifying Ensemble Forecasts, Mon. Weather Rev., 129, 550–560, https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2, 2001.

Hamill, T. M., Whitaker, J. S., and Mullen, S. L.: Reforecasts: An Important Dataset for Improving Weather Predictions, B. Am. Meteorol. Soc., 87, 33–46, https://doi.org/10.1175/BAMS-87-1-33, 2006.

Hamill, T. M., Scheuerer, M., and Bates, G. T.: Analog Probabilistic Precipitation Forecasts Using GEFS Reforecasts and Climatology-Calibrated Precipitation Analyses, Mon. Weather Rev., 143, 3300–3309, https://doi.org/10.1175/MWR-D-15-0004.1, 2015.

Jordan, A., Krüger, F., and Lerch, S.: Evaluating Probabilistic Forecasts with scoringRules, J. Stat. Softw., https://jstatsoft.org, forthcoming, 2018.

Judson, A. and Doesken, N.: Density of Freshly Fallen Snow in the Central Rocky Mountains, B. Am. Meteorol. Soc., 81, 1577–1587, https://doi.org/10.1175/1520-0477(2000)081<1577:DOFFSI>2.3.CO;2, 2000.

Knox, T., Gerhold, L., and Ulbrich, U.: Perception and Use of Uncertainty in Severe Weather Warnings by Emergency Services in Germany, Atmos. Res., 158–159, 292–301, https://doi.org/10.1016/j.atmosres.2014.02.024, 2015.

Lawinenwarndienst Tirol: Winterberichte 2009/2010 bis 2015/2016, available at https://lawine.tirol.gv.at/archiv/winterberichte/ (accessed: 14 November 2017), 2009–2017.

Lerch, S. and Thorarinsdottir, T.: Comparison of Non-Homogeneous Regression Models for Probabilistic Wind Speed Forecasting, Tellus A, 65, 21206, https://doi.org/10.3402/tellusa.v65i0.21206, 2013.

Meister, R.: Density of New Snow and its Dependence on Air Temperature and Wind, Zürcher Geographische Schriften, 23, 73–79, 1985.

Messner, J. W., Mayr, G. J., Wilks, D. S., and Zeileis, A.: Extending Extended Logistic Regression: Extended versus Separate versus Ordered versus Censored, Mon. Weather Rev., 142, 3003–3014, https://doi.org/10.1175/MWR-D-13-00355.1, 2014a.

Messner, J. W., Mayr, G. J., Zeileis, A., and Wilks, D. S.: Heteroscedastic Extended Logistic Regression for Postprocessing of Ensemble Guidance, Mon. Weather Rev., 142, 448–456, https://doi.org/10.1175/MWR-D-13-00271.1, 2014b.

Messner, J. W., Mayr, G. J., and Zeileis, A.: Heteroscedastic Censored and Truncated Regression with crch, The R Journal, 8, 173–181, https://journal.r-project.org/archive/2016-1/messner-mayr-zeileis.pdf (last access: 14 November 2018), 2016.

Mullen, S. L. and Buizza, R.: Quantitative Precipitation Forecasts over the United States by the ECMWF Ensemble Prediction System, Mon. Weather Rev., 129, 638–663, https://doi.org/10.1175/1520-0493(2001)129<0638:QPFOTU>2.0.CO;2, 2001.

Murphy, A. H.: A New Vector Partition of the Probability Score, J. Appl. Meteorol., 12, 595–600, https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2, 1973.

Neal, R. A., Boyle, P., Grahame, N., Mylne, K., and Sharpe, M.: Ensemble Based First Guess Support Towards a Risk-based Severe Weather Warning Service, Meteorol. Appl., 21, 563–577, https://doi.org/10.1002/met.1377, 2014.

Palmer, T. N.: The Economic Value of Ensemble Forecasts as a Tool for Risk Assessment: From Days to Decades, Q. J. Roy. Meteor. Soc., 128, 747–774, https://doi.org/10.1256/0035900021643593, 2002.

Raftery, A. E.: Use and Communication of Probabilistic Forecasts, Stat. Anal. Data Min., 9, 397–410, https://doi.org/10.1002/sam.11302, 2016.

Rasmussen, R., Baker, B., Kochendorfer, J., Meyers, T., Landolt, S., Fischer, A. P., Black, J., Thériault, J. M., Kucera, P., Gochis, D., Smith, C., Nitu, R., Hall, M., Ikeda, K., and Gutmann, E.: How Well Are We Measuring Snow: The NOAA/FAA/NCAR Winter Precipitation Test Bed, B. Am. Meteorol. Soc., 93, 811–829, https://doi.org/10.1175/BAMS-D-11-00052.1, 2012.

Roebber, P. J., Bruening, S. L., Schultz, D. M., and Cortinas Jr., J. V.: Improving Snowfall Forecasting by Diagnosing Snow Density, Weather Forecast., 18, 264–287, https://doi.org/10.1175/1520-0434(2003)018<0264:ISFBDS>2.0.CO;2, 2003.

Rohregger, J. B.: Methoden zur Bestimmung der Schneefallgrenze, Master's thesis, Universität Wien, Austria, 2008.

Roulston, M. S. and Smith, L. A.: Combining Dynamical and Statistical Ensembles, Tellus A, 55, 16–30, https://doi.org/10.1034/j.1600-0870.2003.201378.x, 2003.

Schefzik, R., Thorarinsdottir, T. L., and Gneiting, T.: Uncertainty Quantification in Complex Simulation Models Using Ensemble Copula Coupling, Stat. Sci., 28, 616–640, https://doi.org/10.1214/13-STS443, 2013.

Scheuerer, M.: Probabilistic Quantitative Precipitation Forecasting Using Ensemble Model Output Statistics, Q. J. Roy. Meteor. Soc., 140, 1086–1096, https://doi.org/10.1002/qj.2183, 2014.

Scheuerer, M. and Büermann, L.: Spatially Adaptive Post-Processing of Ensemble Forecasts for Temperature, J. R. Stat. Soc. C-Appl., 63, 405–422, https://doi.org/10.1111/rssc.12040, 2014.

Scheuerer, M. and Hamill, T. M.: Statistical Postprocessing of Ensemble Precipitation Forecasts by Fitting Censored, Shifted Gamma Distributions, Mon. Weather Rev., 143, 4578–4596, https://doi.org/10.1175/MWR-D-15-0061.1, 2015.

Sloughter, J. M. L., Raftery, A. E., Gneiting, T., and Fraley, C.: Probabilistic Quantitative Precipitation Forecasting Using Bayesian Model Averaging, Mon. Weather Rev., 135, 3209–3220, https://doi.org/10.1175/MWR3441.1, 2007.

Stauffer, R., Mayr, G. J., Messner, J. W., Umlauf, N., and Zeileis, A.: Spatio-Temporal Precipitation Climatology over Complex Terrain Using a Censored Additive Regression Model, Int. J. Climatol., 37, 3264–3275, https://doi.org/10.1002/joc.4913, 2017a.

Stauffer, R., Umlauf, N., Messner, J. W., Mayr, G. J., and Zeileis, A.: Ensemble Postprocessing of Daily Precipitation Sums over Complex Terrain Using Censored High-Resolution Standardized Anomalies, Mon. Weather Rev., 145, 955–969, https://doi.org/10.1175/MWR-D-16-0260.1, 2017b.

Thorarinsdottir, T. L. and Gneiting, T.: Probabilistic Forecasts of Wind Speed: Ensemble Model Output Statistics by Using Heteroscedastic Censored Regression, J. R. Stat. Soc. A Stat., 173, 371–388, https://doi.org/10.1111/j.1467-985X.2009.00616.x, 2010.

Umlauf, N., Klein, N., and Zeileis, A.: BAMLSS: Bayesian Additive Models for Location, Scale and Shape (and Beyond), J. Comput. Graph. Stat., 27, 612–627, https://doi.org/10.1080/10618600.2017.1407325, 2017.

Wilks, D. S.: Extending Logistic Regression to Provide Full-Probability-Distribution MOS Forecasts, Meteorol. Appl., 16, 361–368, https://doi.org/10.1002/met.134, 2009.

Zhu, Y., Toth, Z., Wobus, R., Richardson, D., and Myln, E. K.: The Economic Value of Ensemble-Based Weather Forecasts, B. Am. Meteorol. Soc., 83, 73–83, https://doi.org/10.1175/1520-0477(2002)083<0073:TEVOEB>2.3.CO;2, 2002.

Article V

Schlosser L., Hothorn T., Stauffer R., and Zeileis A. (2019). *Distributional Regression Forests for Probabilistic Precipitation Forecasting in Complex Terrain*. Annals of Applied Statistics, 13(3), 1564–1589, doi:10.1214/19-AOAS1247.

JCR ranking: **Top 5** in *Statistics and Probability*.

Contribution (CRT): *Conceptualization / data curation / investigation / validation / writing, review and editing*.

The Annals of Applied Statistics
2019, Vol. 13, No. 3, 1564–1589
https://doi.org/10.1214/19-AOAS1247

DISTRIBUTIONAL REGRESSION FORESTS FOR PROBABILISTIC PRECIPITATION FORECASTING IN COMPLEX TERRAIN

BY LISA SCHLOSSER*, TORSTEN HOTHORN†,[1], RETO STAUFFER* AND ACHIM ZEILEIS*

*Universität Innsbruck** *and Universität Zürich*†

To obtain a probabilistic model for a dependent variable based on some set of explanatory variables, a distributional approach is often adopted where the parameters of the distribution are linked to regressors. In many classical models this only captures the location of the distribution but over the last decade there has been increasing interest in distributional regression approaches modeling all parameters including location, scale and shape. Notably, so-called nonhomogeneous Gaussian regression (NGR) models both mean and variance of a Gaussian response and is particularly popular in weather forecasting. Moreover, generalized additive models for location, scale and shape (GAMLSS) provide a framework where each distribution parameter is modeled separately capturing smooth linear or nonlinear effects. However, when variable selection is required and/or there are nonsmooth dependencies or interactions (especially unknown or of high-order), it is challenging to establish a good GAMLSS. A natural alternative in these situations would be the application of regression trees or random forests but, so far, no general distributional framework is available for these. Therefore, a framework for distributional regression trees and forests is proposed that blends regression trees and random forests with classical distributions from the GAMLSS framework as well as their censored or truncated counterparts. To illustrate these novel approaches in practice, they are employed to obtain probabilistic precipitation forecasts at numerous sites in a mountainous region (Tyrol, Austria) based on a large number of numerical weather prediction quantities. It is shown that the novel distributional regression forests automatically select variables and interactions, performing on par or often even better than GAMLSS specified either through prior meteorological knowledge or a computationally more demanding boosting approach.

1. Introduction. In regression analysis a wide range of models has been developed to describe the relationship between a response variable and a set of covariates. The classical model is the linear model (LM) where the conditional mean of the response is modeled through a linear function of the covariates (see the left panel of Figure 1 for a schematic illustration). Over the last decades this has been extended in various directions including:

Received November 2018; revised February 2019.

[1]An extended research stay of Torsten Hothorn in Innsbruck (August 2017 to January 2018) was financially supported by the Swiss National Science Foundation, grant number SNF IZSEZ0 177091.

Key words and phrases. Parametric models, regression trees, random forests, recursive partitioning, probabilistic forecasting, GAMLSS.

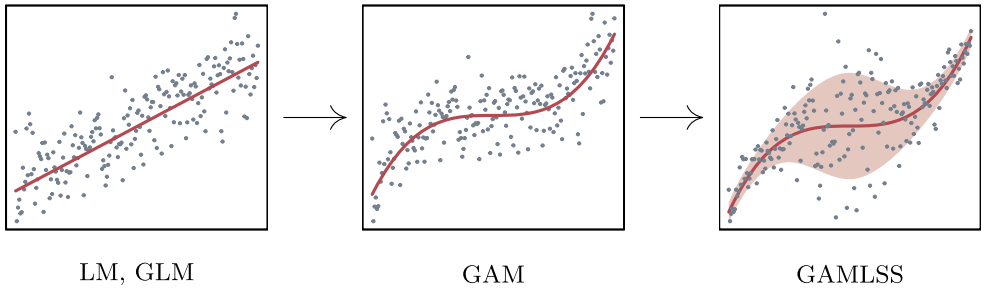

FIG. 1. *Parametric modeling developments*. (*Generalized*) *linear models* (*left*), *generalized additive models* (*middle*), *generalized additive models for location*, *scale and shape* (*right*).

- *Generalized linear models* (GLMs, Nelder and Wedderburn (1972)) encompassing an additional nonlinear link function for the conditional mean.
- *Generalized additive models* (GAMs, Hastie and Tibshirani (1986)) allowing for smooth nonlinear effects in the covariates (Figure 1, middle).
- *Generalized additive models for location, scale and shape* (GAMLSS, Rigby and Stasinopoulos (2005a)) adopting a probabilistic modeling approach. In GAMLSS, each parameter of a statistical distribution can depend on an additive predictor of the covariates comprising linear and/or smooth nonlinear terms (Figure 1, right).

Thus, the above-mentioned models provide a broad toolbox for capturing different aspects of the response (mean only vs. full distribution) and different types of dependencies on the covariates (linear vs. nonlinear additive terms).

While in many applications conditional mean regression models have been receiving the most attention, there has been a paradigm shift over the last decade towards distributional regression models. An important reason for this is that in many fields forecasts of the mean are not the only (or not even the main) concern but instead there is an increasing interest in probabilistic forecasts. Quantities of interest typically include exceedance probabilities for certain thresholds of the response or quantiles of the response distribution. Specifically, consider weather forecasting where there is less interest in the mean amount of precipitation on the next day. Instead, the probability of rain vs. no rain is typically more relevant or, in some situations, a prediction interval of expected precipitation (say from the expected 10% to 90% quantiles). Similar considerations apply for other meteorological quantities and hence attention in the weather forecasting literature has been shifting from classical linear deterministic models (Glahn and Lowry (1972)) towards probabilistic models such as the nonhomogeneous Gaussian regression (NGR) of Gneiting et al. (2005). The NGR typically describes the mean of some meteorological response variable through the average of the corresponding quantity from an ensemble of physically-based numerical weather predictions (NWPs). Similarly, the variance of the response is captured through the variance of the ensemble of NWPs. Thus, the NGR considers both the mean as well as the uncer-

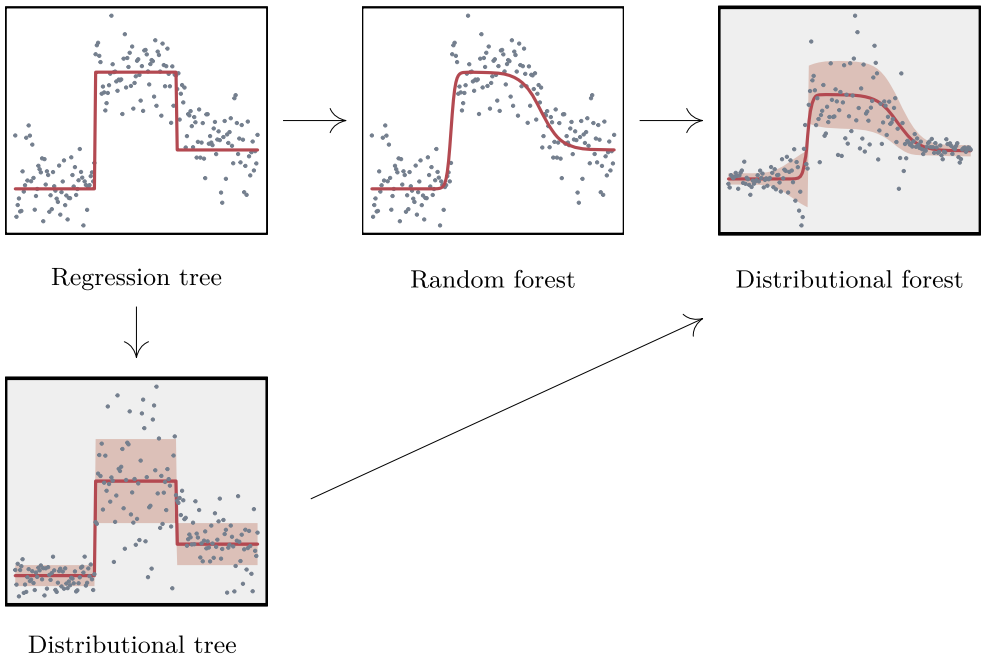

FIG. 2. *Tree and forest developments. Regression tree* (*top left*), *distributional tree* (*bottom left*), *random forest* (*top middle*) *and distributional forest* (*top right*).

tainty of the ensemble predictions to obtain probabilistic forecasts calibrated to a particular site.

In summary, the models discussed so far provide a broad and powerful toolset for parametric distributional fits depending on a specified set of additive linear or smooth nonlinear terms. A rather different approach to capturing the dependence on covariates are tree-based models.

- *Regression trees* (Breiman et al. (1984)) recursively split the data into more homogeneous subgroups and can thus capture abrupt shifts (Figure 2, top left) and approximate nonlinear functions. Furthermore, trees automatically carry out a forward selection of covariates and their interactions.
- *Random forests* (Breiman (2001)) average the predictions of an ensemble of trees fitted to resampled versions of the learning data. This stabilizes the recursive partitions from individual trees and hence better approximates smooth functions (Figure 2, top middle).

While classical regression trees and random forests only model the mean of the response we propose to follow the ideas from GAMLSS modeling—as outlined in Figure 1—and combine tree-based methods with parametric distributional models, yielding two novel techniques:

- *Distributional regression trees* (for short: *distributional trees*) split the data into more homogeneous groups with respect to a parametric distribution, thus captur-

ing changes in any distribution parameter like location, scale or shape (Figure 2, bottom left).

- *Distributional regression forests* (for short: *distributional forests*) utilize an ensemble of distributional trees for obtaining stabilized and smoothed parametric predictions (Figure 2, top right).

In the following, particular focus is given to distributional forests as a method for obtaining probabilistic forecasts by leveraging the strengths of random forests: the ability to capture both smooth and abruptly changing functions along with simultaneous selection of variables and possibly complex interactions. Thus, these properties make the method particularly appealing in case of many covariates with unknown effects and interactions where it would be challenging to specify a distributional regression model like GAMLSS. However, distributional forests should not be considered as a replacement of GAMLSS but rather as a complementing technique for flexible distributional regression—much like GAMs and random forests are complements for conditional mean regression.

In weather forecasting, the flexibility of distributional forests is especially appealing in mountainous regions and complex terrain where a wide range of local-scale effects are not yet resolved by the NWP models. Thus, effects with abrupt changes and possibly nonlinear interactions might be required to account for site-specific unresolved features. To illustrate this in practice, precipitation forecasts are obtained with distributional forests at 95 meteorological stations in a mountainous region in the Alps, covering mainly Tyrol, Austria, and adjacent areas (see the map in Figure 8). More specifically, a Gaussian distribution left-censored at zero, is employed to model 24-hour total precipitation so that the zero-censored point mass describes the probability of observing no precipitation on a given day (see Figure 3). Forecasts for July are established based on data from the same month over the years 1985–2012 including 80 covariates derived from a wide range of different NWP quantities. As Figure 3 shows, the station-wise forests yield a full distributional forecast for each day—here for one specific day (July 24) at one station (Axams) over four years (2009–2012)—based on the previous 24 years as learning data. The corresponding observations conform reasonably well with the predictions. In Section 3 we investigate the performance of distributional forests in this forecasting task in more detail. Compared to three alternative zero-censored Gaussian models distributional forests perform at least on par and sometimes clearly better while requiring no meteorological knowledge about the atmospheric processes which drive formation of precipitation for the model specification. The three alternatives are: a standard ensemble model output statistics approach (EMOS, Gneiting et al. (2005)) based on an NGR, a GAMLSS with regressors prespecified based on meteorological expertise (following Stauffer et al. (2017a)) and a boosted GAMLSS (Hofner, Mayr and Schmid (2016)) using nonhomogeneous boosting (Messner, Mayr and Zeileis (2017)) as an alternative technique for variable selection among all 80 available regressors.

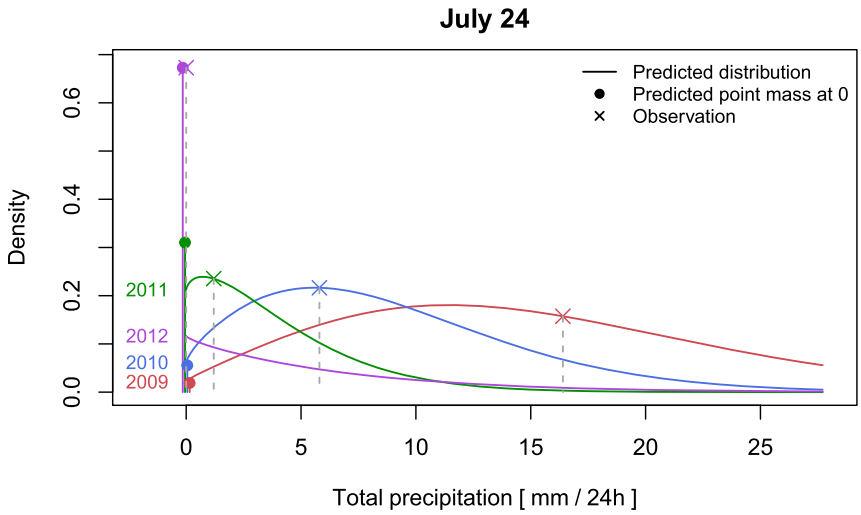

FIG. 3. *Total precipitation predictions by a distributional forest at station Axams for July* 24 *in* 2009, 2010, 2011 *and* 2012 *learned on data from* 1985–2008. *Observations are non-negative and modeled by a Gaussian distribution left-censored at zero. The observations are depicted by crosses and the predicted point mass from the model by filled circles.*

2. Methodology. To embed the distributional approach from GAMLSS into regression trees and random forests, we proceed in three steps. (1) To fix notation, we briefly review fitting distributions using standard maximum likelihood in Section 2.1. (2) A recursive partitioning strategy based on the corresponding scores (or gradients) is introduced in Section 2.2, leading to distributional trees. (3) Ensembles of distributional trees fitted to randomized subsamples are employed to establish distributional forests in Section 2.3.

The general distributional notation is exemplified in all three steps using the Gaussian distribution left-censored at zero (for short: zero-censored Gaussian). The latter is employed in the empirical case study in Section 3 to model power-transformed daily precipitation amounts.

2.1. *Distributional fit.* A distributional model $\mathcal{D}(Y, \boldsymbol{\theta})$ is considered for the response variable $Y \in \mathcal{Y}$ using the distributional family $\mathcal{D}$ with k-dimensional parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and corresponding log-likelihood function $\ell(\boldsymbol{\theta}; Y)$. The GAMLSS framework (Rigby and Stasinopoulos (2005a)) provides a wide range of such distributional families with parameterizations corresponding to location, scale and shape. Furthermore, censoring and/or truncation of these distributions can be incorporated in the usual straightforward way (see e.g., Long (1997, Chapter 7.2)).

To capture both location and scale of the probabilistic precipitation forecasts while accounting for a point mass at zero (i.e., dry days without rain), a zero-censored Gaussian distribution with location parameter μ and scale parameter σ is employed. Therefore, the corresponding log-likelihood function with parameter

vector $\boldsymbol{\theta} = (\mu, \sigma)$ is

$$\ell(\mu, \sigma; Y) = \begin{cases} \log\left\{ \frac{1}{\sigma} \cdot \phi\left(\frac{Y - \mu}{\sigma} \right) \right\} & \text{if } Y > 0, \\ \log\left\{ \Phi\left(\frac{-\mu}{\sigma} \right) \right\} & \text{if } Y = 0, \end{cases} \tag{2.1}$$

where ϕ and Φ are the probability density function and the cumulative distribution function of the standard normal distribution $\mathcal{N}(0, 1)$. Other distributions $\mathcal{D}$ and corresponding log-likelihoods $\ell(\boldsymbol{\theta}; Y)$ could be set up in the same way, for example, for censored shifted gamma distributions (Scheuerer and Hamill (2015)) or zero-censored logistic distributions (Gebetsberger et al. (2017)).

With the specification of the distribution family and its log-likelihood function the task of fitting a distributional model turns into the task of estimating the distribution parameter $\boldsymbol{\theta}$. This is commonly done by maximum likelihood (ML) based on the learning sample with observations $\{y_i\}_{i=1,\ldots,n}$ of the response variable Y. The maximum likelihood estimator (MLE) $\hat{\boldsymbol{\theta}}$ is given by

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{i=1}^{n} \ell(\boldsymbol{\theta}; y_i). \tag{2.2}$$

Equivalently, this can be defined based on the corresponding first-order conditions

$$\sum_{i=1}^{n} s(\hat{\boldsymbol{\theta}}, y_i) = 0, \tag{2.3}$$

where $s(\boldsymbol{\theta}; y_i)$ is the associated score function

$$s(\boldsymbol{\theta}; y_i) = \frac{\partial \ell}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}; y_i). \tag{2.4}$$

The latter is subsequently employed as a general goodness-of-fit measure to assess how well the distribution with parameters $\boldsymbol{\theta}$ fits one individual observation y_i.

2.2. *Distributional tree.* Typically, a single global model $\mathcal{D}(Y, \boldsymbol{\theta})$ is not sufficient for reasonably representing the response distribution. Therefore, covariates $\boldsymbol{Z} = Z_1, \ldots, Z_m \in \mathcal{Z}$ are employed to capture differences in the distribution parameters $\boldsymbol{\theta}$. In weather forecasting, these covariates typically include the output from numerical weather prediction systems and/or lagged meteorological observations.

To incorporate the covariates into the distributional model, they are considered as regressors in additive predictors $g_j(\theta_j) = f_{j,1}(\boldsymbol{Z}) + f_{j,2}(\boldsymbol{Z}) + \cdots$ in GAMLSS. Link functions $g_j(\cdot)$ are used for every parameter θ_j $(j = 1, \ldots, k)$ based on smooth terms $f_{j,k}$ such as nonlinear effects, spatial effects, random coefficients or interaction surfaces (Klein et al. (2015)). However, this requires specifying the additive terms and their functional forms in advance which can be challenging in

practice and potentially require expert knowledge in the application domain, especially if the number of covariates m is large.

Regression trees generally take a different approach for automatically including covariates in a data-driven way and allowing for abrupt changes, nonlinear and nonadditive effects, and interactions. In the context of distributional models the goal is to partition the covariate space $\mathcal{Z}$ recursively into disjoint segments so that a homogeneous distributional model for the response Y can be found for each segment with segment-specific parameters. More specifically, the B disjoint segments $\mathcal{B}_b$ $(b = 1, \ldots, B)$ partition the covariate space

$$\mathcal{Z} = \dot{\bigcup_{b=1,\ldots,B}} \mathcal{B}_b, \tag{2.5}$$

and a local distributional model $\mathcal{D}(Y, \boldsymbol{\theta}^{(b)})$ (i.e., with segment-specific parameters $\boldsymbol{\theta}^{(b)}$) is fitted to the response Y in each segment.

To find the segments $\mathcal{B}_b$ that are (approximately) homogeneous with respect to the distributional model with given parameters, the idea is to use a gradient-based recursive-partitioning approach. In a given subsample of the learning data this fits the model by ML (see equation (2.2)) and then assesses the goodness of fit by assessing the corresponding scores $s(\hat{\boldsymbol{\theta}}; y_i)$ (see equation (2.4)).

To sum up, distributional trees are fitted recursively via:

1. Estimate $\hat{\boldsymbol{\theta}}$ via maximum likelihood for the observations in the current subsample.
2. Test for associations (or instabilities) of the scores $s(\hat{\boldsymbol{\theta}}, y_i)$ and $Z_{l,i}$ for each partitioning variable Z_l $(l = 1, \ldots, m)$.
3. Split the sample along the partitioning variable Z_l^* with the strongest association or instability. Choose the breakpoint with the highest improvement in the log-likelihood or the highest discrepancy.
4. Repeat steps 1–3 recursively in the subsamples until these become too small or there is no significant association/instability (or some other stopping criterion is reached).

Different inference techniques can be used for assessing the association between scores and covariates in step 3. In the following we use the general class of permutation tests introduced by Hothorn et al. (2006) which is also the basis of conditional inference trees (CTree, Hothorn, Hornik and Zeileis (2006)). Alternatively, one could use asymptotic M-fluctuation tests for parameter instability (Zeileis and Hornik (2007)) as in model-based recursive partitioning (MOB, Zeileis, Hothorn and Hornik (2008)). More details are provided in the Appendix.

For obtaining probabilistic predictions from the tree for a (possibly new) set of covariates $\boldsymbol{z} = (z_1, \ldots, z_m)$, the observation simply has to be "sent down" the tree and the corresponding segment-specific MLE has to be obtained. Thus, in practice $\hat{\boldsymbol{\theta}}(\boldsymbol{z})$ does not have to be recalculated for each new $\boldsymbol{z}$ but one can simply

extract the parameter estimates for the corresponding segment which have been computed already while learning the tree. However, to understand this estimator conceptually it is useful to denote it as a weighted MLE where the weights select those observations from the learning sample that fall into the same segment:

$$w_i^{\text{tree}}(z) = \sum_{b=1}^{B} \mathbf{1}\big((z_i \in \mathcal{B}_b) \wedge (z \in \mathcal{B}_b)\big), \tag{2.6}$$

where $\mathbf{1}(\cdot)$ is the indicator function. The predicted distribution for a given z is then fully specified by the estimated parameter $\hat{\boldsymbol{\theta}}(z)$ where

$$\hat{\boldsymbol{\theta}}(z) = \operatorname*{argmax}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{i=1}^{n} w_i^{\text{tree}}(z) \cdot \ell(\boldsymbol{\theta}; y_i). \tag{2.7}$$

2.3. *Distributional forest.* While the simple recursive structure of a tree model is easy to visualize and interpret, the abrupt changes are often too rough, instable, and impose steps on the model even if the true underlying effect is smooth. Hence, ensemble methods such as bagging or random forests (Breiman (2001)) are typically applied to smooth the effects, stabilize the model and improve predictive performance.

The idea of random forests is to learn an ensemble of trees, each on a different learning data obtained through resampling (bootstrap or subsampling). In each node only a random subset of the covariates $\boldsymbol{Z}$ is considered for splitting to reduce the correlation among the trees and to stabilize the variance of the model. For a simple regression random forest the mean of predictions over all trees is considered. In that way changes in the location of the response across the covariates are detected (e.g., in Breiman and Cutler's random forests, Breiman (2001)). This idea is now taken one step further by embedding it in a distributional framework based on maximum-likelihood estimation. *Distributional forests* employ an ensemble of T *distributional trees* which pick up changes in the "direction" of any distribution parameter by considering the full likelihood and corresponding score function for choosing splitting variables and split points.

To obtain probabilistic predictions from a distributional forest, it still needs to be specified how to compute the parameter estimates $\hat{\boldsymbol{\theta}}(z)$ for a (potentially new) set of covariates z. Following Hothorn and Zeileis (2017) we interpret random forests as adaptive local likelihood estimators using the averaged "nearest neighbor weights" (Lin and Jeon (2006)) from the T trees in the forest

$$w_i^{\text{forest}}(z) = \frac{1}{T} \sum_{t=1}^{T} \sum_{b=1}^{B^t} \frac{\mathbf{1}((z_i \in \mathcal{B}_b^t) \wedge (z \in \mathcal{B}_b^t))}{|\mathcal{B}_b^t|}, \tag{2.8}$$

where $|\mathcal{B}_b^t|$ denotes the number of observations in the b-th segment of the t-th tree. Thus, these $w_i^{\text{forest}}(z) \in [0, 1]$ whereas $w_i^{\text{tree}}(z) \in \{0, 1\}$. Hence, weights cannot

only be 0 or 1 but change more smoothly, giving high weight to those observations i from the learning sample that co-occur in the same segment $\mathcal{B}_b^t$ as the new observation z for many of the trees $t = 1, \ldots, T$. Consequently, the parameter estimates may, in principle, change for every observation and can be obtained by

$$\hat{\boldsymbol{\theta}}(z) = \operatorname*{argmax}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{i=1}^{n} w_i^{\text{forest}}(z) \cdot \ell(\boldsymbol{\theta}; y_i). \tag{2.9}$$

In summary, this yields a parametric distributional regression model (through the score-based approach) that can capture both abrupt effects and high-order interactions (through the trees) and smooth effects (through the forest).

Distributional forests share some concepts and algorithmic aspects with other generalizations of Breiman and Cutler's random forests. Nearest neighbor weights are employed for aggregation in survival forests (Hothorn et al. (2004)), quantile regression forests (Meinshausen (2006)), transformation forests (Hothorn and Zeileis (2017)) and generalized random forests for causal inferences (Athey, Tibshirani and Wager (2019)). These procedures aggregate over trees fitted to specific score functions (e.g., log rank scores in survival trees, model residuals in transformation or generalized forests). Distributional forests, in contrast to nonparametric approaches, provide a compromise between model flexibility and interpretability: The parameters of a problem-specific distribution (zero-censored Gaussian for precipitation) have a clear meaning but may depend on external variables in a quite general way.

3. Probabilistic precipitation forecasting in complex terrain. Many statistical weather forecasting models leverage the strengths of modern numerical ensemble prediction systems (EPSs; see Bauer, Thorpe and Brunet (2015)). EPSs not only predict the most likely future state of the atmosphere but provide information about the uncertainty for a specific quantity and weather situation. This is done by running the NWP model several times using slightly perturbed initial conditions and model specifications to account for uncertainties in both, the initial atmospheric state and the NWP model (and its parametrizations). One frequently-used method based on distributional regression models is the ensemble model output statistics (EMOS) approach first proposed by Gneiting et al. (2005) to produce high-quality forecasts for specific quantities and sites. In case of precipitation forecasting, EMOS typically uses the ensemble mean of "total precipitation" (*tp*) forecasts as the predictor for the location parameter μ and the corresponding ensemble standard deviation for the scale parameter σ, for example, assuming the observations to follow a zero-censored Gaussian distribution. This distributional approach of modeling both parameters allows to correct for possible errors of the NWP ensemble in both, the expectation but also the uncertainty of a specific forecast. Thus, a basic EMOS specification typically models the two distribution parameters by two linear predictors, for example, $\mu = \beta_0 + \beta_1 \cdot \text{mean}(tp)$ and

$\log(\sigma) = \gamma_0 + \gamma_1 \cdot \log(\mathrm{sd}(tp))$ with regression coefficients β_0, β_1, γ_0 and γ_1 (where the log link assures positivity of the scale parameter, following Gebetsberger et al. (2017)).

While this approach alone is already highly effective in the plains, it typically does not perform as well in complex terrain due to unresolved effects in the NWP system (Bauer, Thorpe and Brunet (2015)). For example, in the Tyrolean Alps—considered in the following case study—the NWP grid cells of $50 \times 50\,\mathrm{km}^2$ are too coarse to capture single mountains, narrow valleys, etc. Therefore, it is often possible to substantially improve the predictive performance of a basic EMOS by including additional predictor variables, either from local meteorological observations or an NWP model. Unfortunately, it is typically unknown which variables are relevant for improving the predictions. Simply including all available variables may be computationally burdensome and can lead to overfitting but, on the other hand, excluding too many variables may result in a loss of valuable information. Therefore, selecting the relevant variables and interactions among all possible covariates is crucial for improving the statistical forecasting model.

In the following, it is illustrated how distributional forests can solve this problem without requiring prior expert knowledge about the meteorological covariates. For fitting the forest only the response distribution and the list of potential predictor variables need to be specified (along with a few algorithmic details) and then the relevant variables, interactions and potentially nonlinear effects are determined automatically in a data-driven way. Here, we employ a zero-censored Gaussian distribution and 80 predictor variables computed from ensemble means and spreads of various NWP outputs. The predictive performance of the forest is compared to three other zero-censored Gaussian models: (a) a basic EMOS, (b) a GAMLSS with prespecified effects and interactions based on meteorological knowledge/experience, and (c) a boosted GAMLSS with automatic selection of smooth additive terms based on all 80 predictor variables.

3.1. *Data*. Learning and validation data consist of observed daily precipitation sums provided by the National Hydrographical Service (BMLFUW (2016)) and numerical weather forecasts from the U.S. National Oceanic and Atmospheric Administration (NOAA). Both observations and forecasts are available for 1985–2012 and the analysis is exemplified using July, the month with the most precipitation in Tyrol.

Observations are obtained for 95 stations all over Tyrol and surroundings, providing 24-hour precipitation sums measured at 0600 UTC and rigorously quality-checked by the National Hydrographical Service. NWP outputs are obtained from the second-generation reforecast data set of the global ensemble forecast system (GEFS, Hamill et al. (2013)). This data set consists of an 11-member ensemble based on a fixed version of the numerical model and a horizontal grid spacing of about $50 \times 50\,\mathrm{km}^2$ initialized daily at 0000 UTC from December 1984 to present

providing forecasts on a 6-hourly temporal resolution. Each of the 11 ensemble members uses slightly different perturbed initial conditions to predict the situation-specific uncertainty of the atmospheric state.

From the GEFS, 14 basic forecast variables are considered with up to 12 variations each such as mean/maximum/minimum over different aggregation time periods. A detailed overview is provided in Table 1, yielding 80 predictor variables in total.

TABLE 1

Basic covariates together with the number (#) and the type of variations. Time periods indicate aggregation time periods in hours after NWP model initialization (e.g., 6–30 corresponds to +6 h to +30 h ahead forecasts, 0600 UTC to 0600 UTC of the next day)

Basic covariates	#	Variations
tp: total precipitation, power transformed (by 1.6^{-1}) *cape*: convective available potential energy, power transformed (by 1.6^{-1})	12	ensemble mean of sums over 24h, ensemble std. deviation of sums over 24h, ensemble minimum of sums over 24h, ensemble maximum of sums over 24h all for 6–30 ensemble mean of sums over 6h for 6–12, 12–18, 18–24, 24–30 ensemble std. deviation of sums over 6h for 6–12, 12–18, 18–24, 24–30
dswrf: downwards short wave radiation flux ("sunshine") *msl*: mean sea level pressure *pwat*: precipitable water *tmax*: 2m maximum temperature *tcolc*: total column-integrated condensate *t500*: temperature on 500 hPa *t700*: temperature on 700 hPa *t850*: temperature on 850 hPa	6	ensemble mean of mean values, ensemble mean of minimum values*, ensemble mean of maximal values, ensemble std. deviation of mean values, ensemble std. deviation of minimum values*, ensemble std. deviation of maximal values, all over 6–30
tdiff500850: temperature difference 500 to 850 hPa *tdiff500700*: temperature difference 500 to 700 hPa *tdiff700850*: temperature difference 700 to 850 hPa	3	ensemble mean of difference in mean, ensemble minimum of difference in mean, ensemble maximum of difference in mean all over 6–30
msl_diff: mean sea level pressure difference	1	*msl_mean_max* − *msl_mean_min* over 6–30

Note: *Minimum values of *dswrf* over 24 h are always zero and thus neglected.

To remove large parts of the skewness of precipitation data, a power transformation (Box and Cox (1964)) is often applied, for example, using cubic (Stidd (1973)) or square root (Hutchinson (1998)) transformations. However, the power parameter may vary for different climatic zones or temporal aggregation periods and hence we follow Stauffer et al. (2017b) in their choice of 1.6^{-1} as a suitable power parameter for precipitation in the region of Tyrol. The same power transformation is applied to both the observed precipitation sums and the NWP outputs "total precipitation" (*tp*) and "convective available potential energy" (*cape*).

3.2. *Models and evaluation.* The following zero-censored Gaussian regression models are employed in the empirical case study, see Table 2 for further details:

- *Distributional forest*: All 80 predictor variables are considered for learning a forest of 100 trees. Subsampling is employed for each tree using a third of the predictors in each split of the tree (argument `mtry` in our implementation `distforest`, with more "computational details" provided at the end of the manuscript). Parameters are estimated by adaptive local likelihood based on the forest weights as described in Section 2. The stopping criteria are the minimal number of observations to perform a split (`minsplit` $= 50$), the minimal number of observations in a segment (`minbucket` $= 20$) and the significance level

TABLE 2

Overview of models with type of covariate dependency and included covariates for each distribution parameter. A $*$ B *indicates an interaction between covariate* A *and* B

Model	Type	Location (μ)	Scale ($\log(\sigma)$)
Distributional forest	recursive partitioning	all	all
EMOS	linear	*tp_mean*	*tp_sprd*
Prespecified GAMLSS	spline in each	*tp_mean*, *tp_max*, *tp_mean1218* $*$ *cape_mean1218*, *dswrf_mean_mean*, *tcolc_mean_mean*, *pwat_mean_mean*, *tdiff500850_mean*, *msl_diff*	*tp_sprd*, *dswrf_sprd_mean*, *tp_sprd1218* $*$ *cape_mean1218*, *tcolc_sprd_mean*, *tdiff500850_mean*
Boosted GAMLSS	spline in each	all	all

for variable selection (`alpha` $= 1$). The latter means that no early stopping (or "prepruning") is applied based on results of the statistical tests.

- *EMOS*: The basic ensemble model output statistics models use the ensemble mean of total precipitation as regressor in the location submodel and the corresponding ensemble standard deviation in the scale submodel. The parameters are estimated by maximum likelihood, using an identity link for the location part and a log link for the scale part (following the advice of Gebetsberger et al. (2017)).
- *Prespecified GAMLSS*: Smooth additive splines are selected for the most relevant predictors based on meteorological expert knowledge following Stauffer et al. (2017a). More specifically, based on the 80 available variables, eight terms are included in the location submodel and five in the scale submodel. Both involve an interaction of *tp* and *cape* in the afternoon (between 1200 UTC and 1800 UTC) to capture the potential for thunderstorms that frequently occur in summer afternoons in the Alps. The model is estimated by maximum penalized likelihood using a backfitting algorithm (Stasinopoulos and Rigby (2007)).
- *Boosted GAMLSS*: Smooth additive splines are selected automatically from all 80 available variables, using noncyclic boosting for parameter estimation (Hofner, Mayr and Schmid (2016), Messner, Mayr and Zeileis (2017)). This updates the predictor terms for the location or scale submodels iteratively by maximizing the log-likelihood only for the variable yielding the biggest improvement. The iteration stops early—before fully maximizing the in-sample likelihood—based on a (computationally intensive) out-of-bag bootstrap estimate of the log-likelihood. The grid considered for the number of boosting iterations (`mstop`) is: $50, 75, \ldots, 975, 1000$.

The predictive performance in terms of full probabilistic forecasts is assessed using the continuous ranked probability score (CRPS, Hersbach (2000)). For each of the models this assesses the discrepancy of the predicted distribution function F from the observation y by

$$\text{CRPS}(y, F) = \int_{-\infty}^{\infty} \big(F(z) - \mathbf{1}(y \le z)\big)^2 \, dz \tag{3.1}$$

where $\mathbf{1}(\cdot)$ is the indicator function. In the subsequent applications, the mean CRPS is always evaluated out of sample, either using cross-validation or a hold-out data set (2009–2012) that was not used for learning (1985–2008). CRPS is a proper scoring rule (Gneiting and Raftery (2007)) often used within the meteorological community. Lower values indicate better performance.

To assess differences in the improvement of the forests and GAMLSS models over the basic EMOS, a CRPS-based skill score with EMOS as the reference method is computed:

$$\text{CRPSS}_{\text{method}} = 1 - \frac{\text{CRPS}_{\text{method}}}{\text{CRPS}_{\text{EMOS}}}. \tag{3.2}$$

3.3. *Application for one station.* In a first step, we show a detailed comparison of the competing models for one observation site, Axams in Tyrol (in the center of the study area, see Figure 8). As for all other stations, daily precipitation observations and numerical weather predictions are available for the month of July from 1985 through 2012. In Figure 3 in the introduction the probabilistic forecasts from the distributional forest, trained on 1985–2008, for July 24 in 2009–2012 have already been shown as a motivational example. In particular, the figure depicts the forecasted point mass at zero (i.e., the probability of a dry day) along with the forecasted probability density function for the total amount of precipitation. Based on this illustration it can be observed that the four sample forecasts differ considerably in location μ, scale σ, and the amount of censoring while conforming quite well with the actual observations from these days. While this is a nice illustrative example we are interested in the overall predictive performance and calibration of the distributional fits. More details of this assessment as well as an application to 14 further meteorological stations is provided in Supplement B (Schlosser et al. (2019b)).

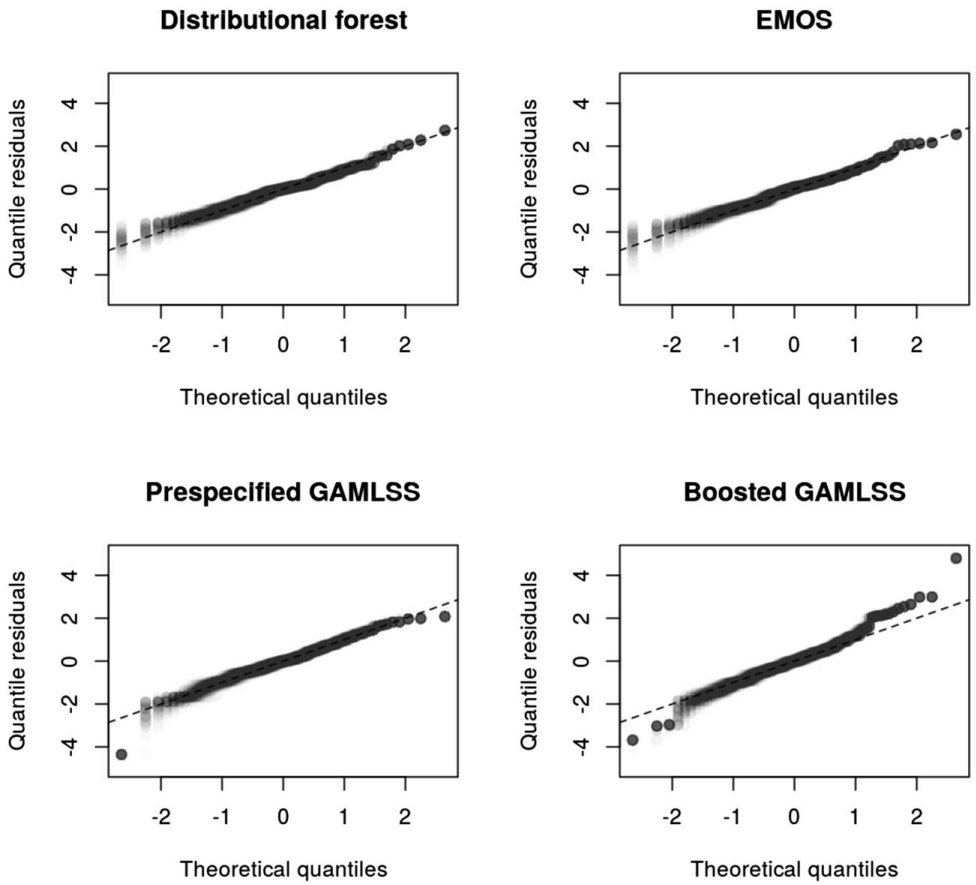

FIG. 4. *Out-of-sample residual QQ plots* (2009–2012) *for station Axams based on models learned on data from* 1985–2008.

To assess calibration, Figure 4 shows residual QQ plots for out-of-sample predictions (2009–2012) from the different models trained on 1985–2008. Due to the point masses at zero 100 draws from the randomized quantile residuals (Dunn and Smyth (1996)) are plotted in semi-transparent gray. Overall, the randomized quantile residuals conform quite well with the theoretical standard normal quantile (i.e., form a straight line close to the diagonal), indicating that all four models are sufficiently well calibrated. This is also supported by the corresponding probability integral transform (PIT, Gneiting, Balabdaoui and Raftery (2007)) histograms for station Axams in Supplement B (Schlosser et al. (2019b)) which contains a more detailed explanation of residual QQ plots and PIT histograms.

To assess the predictive performance, a full cross-validation is carried out rather than relying on just the one fixed test set for the years 2009–2012. To do so, a 10 times 7-fold cross-validation is carried out where each repetition splits the available 28 years into 7 subsets of 4 randomly-selected (and thus not necessarily consecutive) years. The models are learned on 6 folds (= 24 years) and evaluated on the 7th fold (= 4 years) using the average CRPS across all observations. The resulting 10 CRPS skill scores are displayed by boxplots in Figure 5 using EMOS as the reference model (horizontal line at a CRPSS of 0). Both GAMLSS models and the distributional forest perform distinctly better than the EMOS model. While the two GAMLSS lead to an improvement of around 4 percent, the distributional forest has a slightly higher improvement of around 5.5 percent in median.

Finally, it is of interest how this improvement in predictive performance by the distributional forest is accomplished, that is, which of the 80 covariates are selected in the trees of the forest. As the 100 trees of the forest do not allow to simply assess the variables' role graphically, a common solution for random forests in

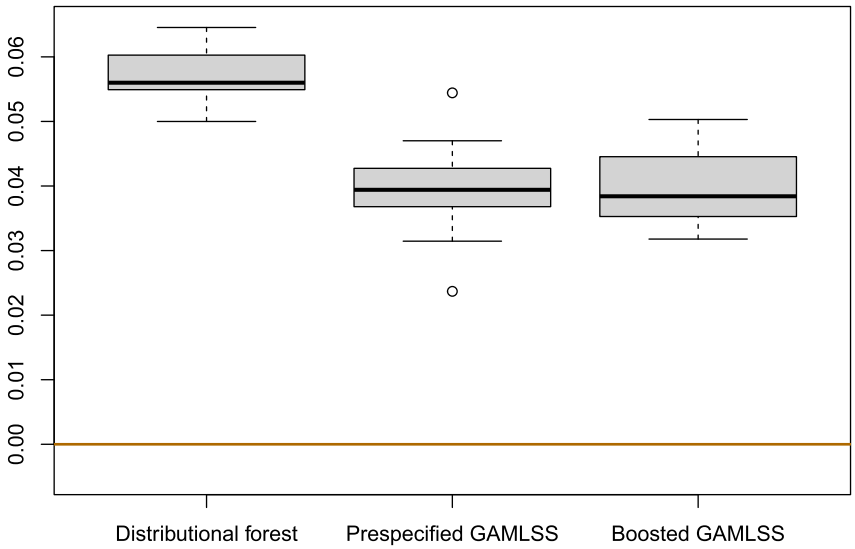

FIG. 5. *CRPS skill score from the* 10 *times* 7*-fold cross-validation at station Axams* (1985–2012). *The horizontal orange line pertains to the reference model EMOS.*

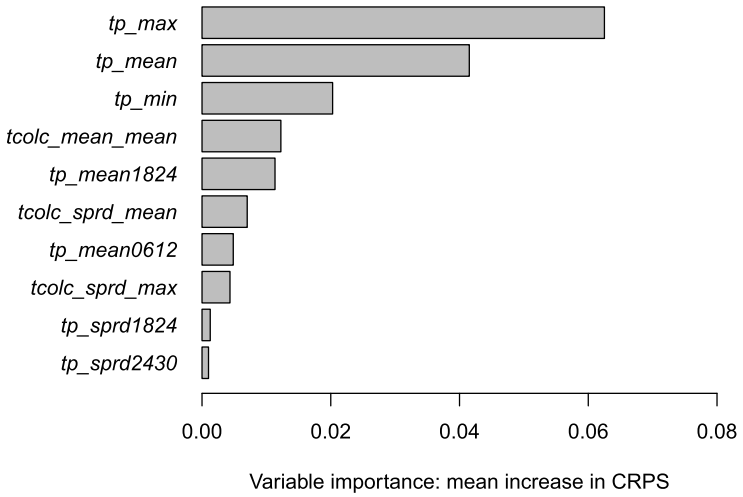

FIG. 6. *CRPS-based variable importance for the top* 10 *covariates in the distributional forest. Based on data for station Axams, learning period* 1985–2008 *and assessed in* 2009–2012.

general is to consider variable importance measures. Here, this is defined as the amount of change in CRPS when the association between one covariate and the response variable is artificially broken through permutation (and thus also breaking the association to the remaining covariates).

Figure 6 shows the 10 covariates with the highest permutation importance (i.e., change in CRPS) for station Axams. As expected the NWP outputs for total precipitation (*tp*) are particularly important along with total column-integrated condensate (*tcolc*). Also, both variables occur in various transformations such as means (either of the full day or certain parts of the afternoon), spreads or minima/maxima. Thus, while the covariates themselves are not surprising, selecting a GAMLSS with a particular combination of all the transformations would be much more challenging.

3.4. *Application for all stations.* After considering only one observational site up to now, a second step evaluates and compares the competing methods on all 95 available stations. As in the previous section, all models are learned on the first 24 years and evaluated by the average CRPS on the last 4 years. More specifically, the CRPS skill score against the EMOS model is computed for the out-of-sample predictions at each station and visualized by parallel coordinates plots with boxplots superimposed in Figure 7. Overall, distributional forests have a slightly higher improvement in CRPSS compared to the two GAMLSS which is best seen by looking at the boxplots and the green line representing the results for station Axams. The underlying parallel coordinates additionally bring out that the prespecified GAMLSS sometimes performs rather differently (sometimes better, sometimes worse) compared to the two data-driven models. Values below zero show that, for some stations, EMOS performs better than the more complex statistical methods.

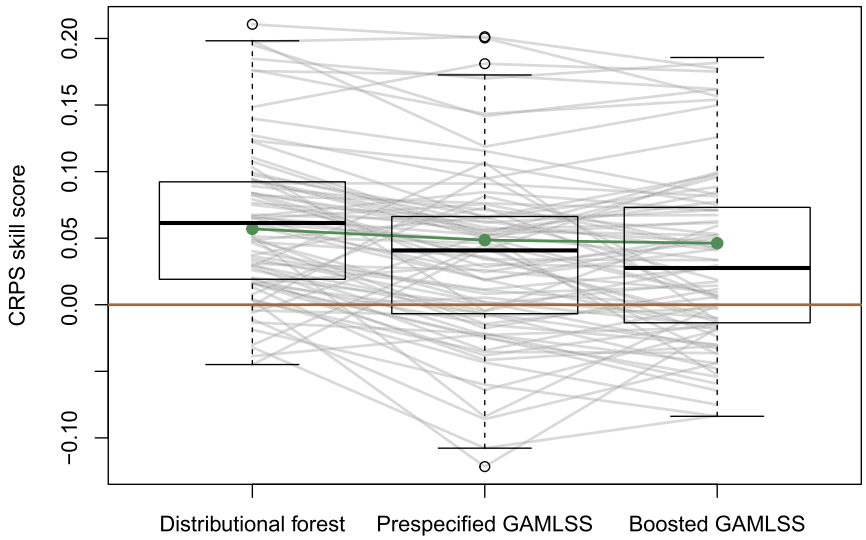

FIG. 7. *CRPS skill score for each station* (*gray lines with boxplots superimposed*). *Station Axams is highlighted in green and the horizontal orange line pertains to the reference model EMOS. The models are learned on* 1985–2008 *and validated for* 2009–2012.

To assess whether these differences in predictive performance are due to differences in the topography, Figure 8 shows a brief spatial summary of all stations. Each station is illustrated by a symbol that conveys which model performed best in terms of CRPS on the last four years of the data. Additionally, the color of the symbol indicates the CRPS difference between distributional forest and the best-performing other model. Green signals that the distributional forest performs better than the other models whereas red signals that another model performs better. Overall the distributional forest performs on par (gray) or better (green) for the majority of stations. Only for a few stations in the north-east EMOS performs best, and in East Tyrol the prespecified GAMLSS performs particularly well in the validation period (2009–2012). Partially, this can be attributed to random variation as the differences at several stations are mitigated when considering a full cross-validation rather than a single split into learning and validation period (see Supplement B, Schlosser et al. (2019b) and the corresponding discussion in the next section). Further differences are possibly due to East Tyrol lying in a different climate zone, south of the main Alpine Ridge. Hence, long-term climatological characteristics as well as the precipitation patterns in 2009–2012 differ from North Tyrol, conforming particularly well with the additive effects from the prespecified GAMLSS.

4. Discussion. Distributional regression modeling is combined with tree-based modeling to obtain a novel and flexible method for probabilistic forecasting. The resulting distributional trees and forests can capture abrupt and nonlinear effects and interactions in a data-driven way. By basing the split point and split variable selection on a full likelihood and corresponding score function, the trees

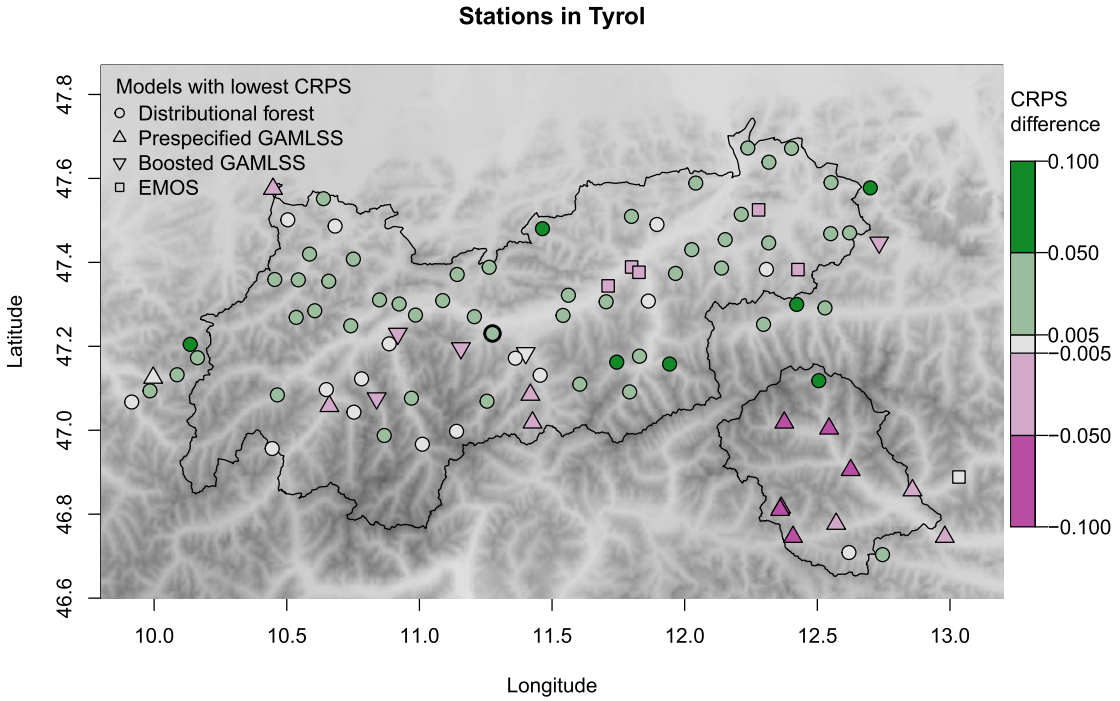

FIG. 8. *Map of Tyrol coding the best-performing model for each station (type of symbol) when learned on* 1985–2008 *and validated for* 2009–2012. *The color codes whether the distributional forest had higher (green) or lower (red) CRPS compared to the best of the other three models. The gray background shows the local topography (Robinson, Regetz and Guralnick* (2014)). *Station Axams is highlighted in bold.*

and forests can not only pick up changes in the location but also the scale or shape of any distributional family.

Distributional forests are an attractive alternative when prespecifying or boosting all possible effects and interactions in a GAMLSS model is challenging. Distributional forests are rather straightforward to specify requiring only little prior subject matter knowledge and also work well in the presence of many potential covariates. The application to precipitation forecasting in complex terrain illustrates that distributional forests often perform on par or even better than their GAMLSS counterparts. Hence, they form a useful addition to the already available toolbox of probabilistic forecasts for disciplines such as meteorology.

Variable selection. Generally, there are many possibilities how to specify the variables that are to be included in a distributional regression model. Especially for a low number of covariates, the GAMLSS approach offers a powerful framework in which penalized estimation of both smooth main effects and corresponding interaction surfaces yields models that often balance good predictive performance with high interpretability (see e.g., Wood, Scheipl and Faraway (2013); Goicoa et al. (2018); Ugarte, Adin and Goicoa (2017b)). However, if the number of covariates is high, including all (or many) main effects and interactions in a GAMLSS

typically becomes challenging both in terms of interpretability and computational complexity/stability (see also Hofner, Mayr and Schmid (2016)).

In the precipitation forecasting application, as presented in Section 3, 80 covariates are considered which corresponds to 3160 potential pairwise interactions (and even more higher-order interactions). Therefore, only main effects are considered for the boosted GAMLSS while the prespecified GAMLSS also includes selected interactions chosen based on meteorological expert knowledge. In contrast, the distributional forest requires no prespecification as covariates and corresponding interactions are selected automatically. Thus, distributional forests are an appealing alternative to (boosted) GAMLSS in weather forecasting tasks as the main concern is typically not so much interpretability but forecasting skill and (semi-)automatic application on a larger domain (see also the discussion in Rasp and Lerch (2018)).

Distributional specifications for precipitation modeling. Choosing an adequate distributional family is an important step for establishing a well-fitting model. A zero-censored Gaussian distribution is employed in this manuscript as this has been found to be an appropriate choice for precipitation modeling in earlier literature (e.g., Stauffer et al. (2017b)). To test for robustness against distributional misspecification, two alternative distributional specifications have been considered in Supplement A (Schlosser et al. (2019a)): Using the same evaluations as in Section 3.4, all models are additionally fitted for 15 meteorological stations using a zero-censored logistic distribution in order to account for heavier tails and a two-part Gaussian hurdle model combining a binary model for zero vs. positive precipitation and a separate Gaussian model, truncated at zero, for the positive precipitation observations. Both specifications yield qualitatively similar results as for the zero-censored Gaussian distribution. For some stations the two-part hurdle model leads to small improvements, however at the expense of increased variability across stations (especially for EMOS and the boosted GAMLSS). Overall, the results from this manuscript are quite robust across these distributional specifications, especially for the distributional forests.

Moreover, one could consider a distribution including an additional parameter for capturing skewness (as in Scheuerer and Hamill (2015), Baran and Nemoda (2016)). However, this would go beyond the mean/variance specification of the NGR that is widely used in ensemble post-processing. Therefore, this contribution investigates the effects of using the same distributional family with a novel strategy for specifying dependence on covariates.

More general distributional specifications. Beyond the task of modeling precipitation it is of interest how well distributional forests perform in combination with other more general distributional specifications. It has been shown previously in the literature that using a score- or gradient-based selection of splitting variables outperforms a mean-based selection with subsequent flexible distributional modeling: For example, both Athey, Tibshirani and Wager (2019, Figure 2) and

Hothorn and Zeileis (2017, Figure 1) demonstrated (independently) that their respective score-based random forest algorithms outperform the mean-based quantile regression forests of Meinshausen (2006) in a setup where only the variance of a normal response variable changes across the considered covariates. However, if all distribution parameters are closely correlated with the distribution mean the forests with different splitting strategies all perform similarly, provided a sufficiently flexible distribution is employed for the final predictions (see Hothorn and Zeileis (2017), Section 7).

Similarly, the score-based distributional forests introduced in this manuscript proved to be quite robust to the different distributional specifications considered. While all specifications focus on capturing mean-variance effects note that these parameters are never fully orthogonal but can actually become quite closely correlated due to the censoring (or truncation and/or zero-inflation considered in Supplement A (Schlosser et al. (2019a))).

However, exploring extensions to more flexible parametric distributions (e.g., such as the Dagum distribution considered by Klein et al. (2015) in GAMLSS-type models) as well as transformation model specifications (e.g., as in Hothorn and Zeileis (2017)) are of interest for future research.

Axams vs. other meteorological stations. Axams was chosen as the meteorological station for the more extensive evaluations in Section 3.3 as it yields fairly typical results and is geographically in the center of the study area and closest to Innsbruck, the capital of Tyrol and the work place of three of the authors. To show that qualitatively similar results are obtained for other meteorological stations, Supplement B (Schlosser et al. (2019b)) carries out the same evaluation for 14 further stations. These cover a wide range of geographical locations/altitudes and a mix of different best-performing models in the single-split setting reported in Section 3.4.

The supplement shows that some of the differences in forecast skill from Figure 8 even out in the cross-validation with distributional forests typically performing at least as well as the best of the other models at most stations. In particular, this also includes three stations in East Tyrol where the prespecified GAMLSS performs best in the single-split setting (learning based on 1985–2008 and validation for 2009–2012).

Tuning parameters. Selecting tuning parameters for flexible regression models is important not only in terms of predictive accuracy but also computational complexity. For the application in Section 3 tuning parameters are selected based on advice from the literature as well as our own experiences. As Hastie, Tibshirani and Friedman (2001) and Breiman (2001) recommend to build full-grown trees, early stopping upon nonsignificance is disabled (`alpha` $= 1$) and low values are used for `minsplit` ($= 50$) and `minbucket` ($= 20$), while assuring that

`minsplit` is sufficiently large for reasonably obtaining MLEs of all parameters in each segment of the tree.

Applying the Law of Large Numbers it can be shown that random forests do not overfit as the number of trees increases (Biau and Scornet (2016), Breiman (2001), Hastie, Tibshirani and Friedman (2001)). Therefore, in principle, forests can be built with a very large number of trees (`ntree`) as this cannot deteriorate the predictions. However, "[...] the computational cost for inducing a forest increases linearly with the number of trees, so a good choice results from a trade-off between computational complexity and accuracy" (Biau and Scornet (2016, p. 205)). Following this advice, we decided to build forests consisting of 100 trees.

Computational difficulties. As stated by Hofner, Mayr and Schmid (2016) the AIC-based variable selection methods implemented in the R package *gamlss* "[...] can be unstable, especially when it comes to selecting possibly different sets of variables for multiple distribution parameters." We have noticed computational problems when applying *gamlss* in certain settings within the cross-validation framework as it did not succeed in fitting the model. In these cases the prespecified GAMLSS was not taken into consideration in the comparison of all applied models.

Computational details. The proposed methods are implemented in the R package *disttree* (version 0.1.0) based on the *partykit* package (version 1.2.3), both available on R-Forge at (https://R-Forge.R-project.org/projects/partykit/). The function `distforest` learns the distributional forests proposed in this manuscript by combining the general `cforest` function from *partykit* with the function `distfit` for fitting distributional models by maximum likelihood. Analogously, `disttree` can learn a single distributional tree by combining `ctree` with `distfit`. All functions can either be used with GAMLSS family objects from the R package *gamlss.dist* (Stasinopoulos and Rigby (2007), version 5.0.6) or with custom lists containing all required information about the distribution family.

In addition to *disttree*, Section 3 employs R package *crch* (Messner, Mayr and Zeileis (2016), version 1.0.1) for the EMOS models, *gamlss* (Stasinopoulos and Rigby (2007), version 5.1.0) for the prespecified GAMLSS and *gamboostLSS* (Hofner, Mayr and Schmid (2016), version 2.0.1) for the boosted GAMLSS.

The fitted distributional forest for July 24 and observation station Axams (including Figure 3) is reproducible using `demo("RainAxams", package = "disttree")`. This also includes fitting the other zero-censored Gaussian models considered in this paper and generating the corresponding QQ plots (Figure 4) and PIT histograms (Schlosser et al. (2019b), Supplement B). Full replication of all results can be obtained with `demo("RainTyrol", package = "disttree")` requiring the companion R package *RainTyrol* (version 0.1.0), also available within the R-Forge project. The results presented

in Supplement A (Schlosser et al. (2019a)) and Supplement B (Schlosser et al. (2019b)) can be reproduced using `demo("RainDistributions", package = "disttree")` and `demo("RainStationwise", package = "disttree")`, respectively.

APPENDIX: TREE ALGORITHM

In the following, the tree algorithm applied in the empirical case study discussed in this paper is explained. For notational simplicity, the testing and splitting procedure is described for the root node, that is, the entire learning sample with observations $\{y_i\}_{i=1,\dots,n}$, $n \in \mathbb{N}$. In each child node the corresponding subsample depends on the foregoing split(s).

After fitting a distributional model $\mathcal{D}(Y, \boldsymbol{\theta})$ to the learning sample with observations $\{y_i\}_{i=1,\dots,n}$ as explained in Section 2.1 the resulting estimated parameter $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \dots, \hat{\theta}_k)$, $k \in \mathbb{N}$ can be plugged in the score function $s(\boldsymbol{\theta}, Y)$. In that way a goodness-of-fit measurement is obtained for each parameter θ_j and each observation y_i. To use this information, statistical tests are employed to detect dependencies between the score values

$$s(\hat{\boldsymbol{\theta}}, y) = \begin{pmatrix} s(\hat{\boldsymbol{\theta}}, y_1)_1 & s(\hat{\boldsymbol{\theta}}, y_1)_2 & \dots & s(\hat{\boldsymbol{\theta}}, y_1)_k \\ \vdots & \vdots & \ddots & \vdots \\ s(\hat{\boldsymbol{\theta}}, y_n)_1 & s(\hat{\boldsymbol{\theta}}, y_n)_2 & \dots & s(\hat{\boldsymbol{\theta}}, y_n)_k \end{pmatrix} \tag{A.1}$$

and each variable $Z_l \in \{Z_1, \dots, Z_m\}$. More formally, the following hypotheses are assessed with permutation tests:

$$H_0^l : s(\hat{\boldsymbol{\theta}}, Y) \quad \perp \quad Z_l. \tag{A.2}$$

The permutation tests are based on the multivariate linear statistic

$$T_l = \operatorname{vec}\left(\sum_{i=1}^{n} v_l(Z_{li}) \cdot s(\hat{\boldsymbol{\theta}}, Y_i) \right), \tag{A.3}$$

where $s(\hat{\boldsymbol{\theta}}, Y_i) \in \mathbb{R}^{1 \times k}$ and the type of the transformation function v_l depends on the type of the split variable Z_l. If Z_l is numeric then v_l is simply the identity function $v_l(Z_{li}) = Z_{li}$ and therefore $T_l \in \mathbb{R}^k$ as the "vec" operator converts the $1 \times k$ matrix into a k column vector. If Z_l is a categorical variable with H categories then $v_l(Z_{li}) = (\mathbf{I}(Z_{li} = 1), \dots, \mathbf{I}(Z_{li} = H))$ such that v_l is a H-dimensional unit vector where the element corresponding to the value of Z_{li} is 1. In this case the statistic $T_l \in \mathbb{R}^{H \cdot k}$ as the "vec" operator converts the $H \times k$ matrix into a $H \cdot k$ column vector by column-wise combination. Observations with missing values are excluded from the sums.

With the conditional expectation μ_l and the covariance Σ_l of T_l as derived by Strasser and Weber (1999) the test statistic can be standardized. The observed multivariate linear statistic t_l which is either a k- or $k \cdot H$-dimensional vector, depending on the scale of Z_l, is mapped onto the real line by a univariate test statistic c.

In the application of this paper a quadratic form is chosen, such that

$$c_{\text{quad}}(t_l, \mu_l, \Sigma_l) = (t_l - \mu_l)\Sigma_l^{+}(t_l - \mu_l)^{\top} \tag{A.4}$$

where Σ_l^{+} is the Moore–Penrose inverse of Σ_l. Alternatively, the maximum of the absolute values of the standardized linear statistic can be considered ($c_{\max}$).

Strasser and Weber (1999) showed that the asymptotic conditional distribution of the linear statistic t_l is a multivariate normal with parameters μ_l and Σ_l. Hence, the asymptotic conditional distribution of $c(t_l, \mu_l, \Sigma_l)$ is either normal (for $c_{\max}$) or χ^2 (for c_{quad}).

The smaller the p-value corresponding to the standardized test statistic $c(t_l, \mu_l, \Sigma_l)$ is the stronger the discrepancy from the assumption of independence between the scores and the split variable Z_l. After Bonferroni-adjusting the p-values it has to be assessed whether any of the resulting p-values are beneath the selected significance level. If so, the partitioning variable Z_{l^*} with the lowest p-value is chosen as splitting variable. Otherwise no further split is made in this node as the stopping criterion of no p-values being below the significance level is fulfilled. This type of early-stopping in building a tree is sometimes also referred to as "prepruning". For random forests prepruning is often switched off by setting the significance level to 1.

The breakpoint that leads to the highest discrepancy between score functions in the two resulting subgroups is selected as split point. This is measured by the linear statistic

$$T_{l^*}^{qr} = \sum_{i \in \mathcal{B}_{qr}} s(\hat{\boldsymbol{\theta}}, Y_i) \tag{A.5}$$

for $q \in \{1, 2\}$ where $\mathcal{B}_{1r}$ and $\mathcal{B}_{2r}$ are the two new subgroups, without any particular ordering, that are defined by splitting in split point r of variable Z_{l^*}. The split point is then chosen as follows:

$$r^* = \operatorname*{argmin}_r \Big(\min_{q=1,2} \big(c\big(t_{l^*}^{qr}, \mu_{l^*}^{qr}, \Sigma_{l^*}^{qr}\big)\big)\Big). \tag{A.6}$$

One repeats the testing and splitting procedure in each of the resulting subgroups until some stopping criterion is reached. This criterion can for example be a minimal number of observations in a node or a minimal p-value for the statistical tests. In that way prepruning is applied in order to find right-sized trees and hence avoid overfitting.

This permutation-test-based tree algorithm is presented in Hothorn, Hornik and Zeileis (2006) as the CTree algorithm. A different framework to build a likelihood-based tree is provided by the MOB algorithm which is based on M-fluctuation tests (Zeileis, Hothorn and Hornik (2008)).

SUPPLEMENTARY MATERIAL

Supplement A: Different response distributions (DOI: 10.1214/19-AOAS1247SUPPA; .pdf). To assess the goodness of fit of the Gaussian distribution, left-censored at zero, this supplement employs the same evaluations as in the main manuscript but based on two other distributional assumptions: A logistic distribution, left-censored at zero, is employed to potentially better capture heavy tails—and a two-part hurdle model combining a binary model for zero vs. positive precipitation and a Gaussian model, truncated at zero, for the positive precipitation observations.

Supplement B: Stationwise evaluation (DOI: 10.1214/19-AOAS1247SUPPB; .pdf). To show that Axams is a fairly typical station and similar insights can be obtained for other stations as well, this supplement presents the same analysis as in Section 3.3 of the main manuscript for 14 further meteorological stations.

REFERENCES

ATHEY, S., TIBSHIRANI, J. and WAGER, S. (2019). Generalized random forests. *Ann. Statist.* **47** 1148–1178. MR3909963

BARAN, S. and NEMODA, D. (2016). Censored and shifted gamma distribution based EMOS model for probabilistic quantitative precipitation forecasting. *Environmetrics* **27** 280–292. MR3521202

BAUER, P., THORPE, A. and BRUNET, G. (2015). The quiet revolution of numerical weather prediction. *Nature* **525** (7567) 47–55.

BIAU, G. and SCORNET, E. (2016). A random forest guided tour. *TEST* **25** 197–227. MR3493512

BMLFUW (2016). Bundesministerium für Land und Forstwirtschaft, Umwelt und Wasserwirtschaft (BMLFUW), Abteilung IV/4—Wasserhaushalt. Available at http://ehyd.gv.at/. Accessed: 2016–02–29.

BOX, G. E. P. and COX, D. R. (1964). An analysis of transformations. (With discussion). *J. Roy. Statist. Soc. Ser. B* **26** 211–252. MR0192611

BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees. Wadsworth Statistics/Probability Series*. Wadsworth Advanced Books and Software, Belmont, CA. MR0726392

DUNN, P. K. and SMYTH, G. K. (1996). Randomized quantile residuals. *J. Comput. Graph. Statist.* **5** 236–244.

GEBETSBERGER, M., MESSNER, J. W., MAYR, G. J. and ZEILEIS, A. (2017). Fine-tuning nonhomogeneous regression for probabilistic precipitation forecasts: Unanimous predictions, heavy tails, and link functions. *Mon. Weather Rev.* **145** 4693–4708.

GLAHN, H. R. and LOWRY, D. A. (1972). The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteorol.* **11** 1203–1211.

GNEITING, T., BALABDAOUI, F. and RAFTERY, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 243–268. MR2325275

GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. MR2345548

GNEITING, T., RAFTERY, A. E., WESTVELD III, A. H. and GOLDMAN, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Weather Rev.* **133** 1098–1118.

GOICOA, T., ADIN, A., UGARTE, M. D. and HODGES, J. S. (2018). In spatio-temporal disease mapping models, identifiability constraints affect PQL and INLA results. *Stoch. Environ. Res. Risk Assess.* **32** 749–770.

HAMILL, T. M., BATES, G. T., WHITAKER, J. S., MURRAY, D. R., FIORINO, M., GALARNEAU JR., T. J., ZHU, Y. and LAPENTA, W. (2013). NOAA's second-generation global medium-range ensemble reforecast dataset. *Bull. Am. Meteorol. Soc.* **94** 1553–1565.

HASTIE, T. and TIBSHIRANI, R. (1986). Generalized additive models. *Statist. Sci.* **1** 297–318. MR0858512

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*. *Springer Series in Statistics*. Springer, New York. MR1851606

HERSBACH, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather Forecast.* **15** 559–570.

HOFNER, B., MAYR, A. and SCHMID, M. (2016). gamboostLSS: An R package for model building and variable selection in the GAMLSS framework. *J. Stat. Softw.* **74** (1) 1–31.

HOTHORN, T., HORNIK, K. and ZEILEIS, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *J. Comput. Graph. Statist.* **15** 651–674. MR2291267

HOTHORN, T. and ZEILEIS, A. (2017). Transformation forests. Available at arXiv:1701.02110.

HOTHORN, T., LAUSEN, B., BENNER, A. and RADESPIEL-TRÖGER, M. (2004). Bagging survival trees. *Stat. Med.* **23** 77–91.

HOTHORN, T., HORNIK, K., VAN DE WIEL, M. A. and ZEILEIS, A. (2006). A Lego system for conditional inference. *Amer. Statist.* **60** 257–263. MR2246759

HUTCHINSON, M. F. (1998). Interpolation of rainfall data with thin plate smoothing splines—Part II: Analysis of topographic dependence. *Journal of Geographic Information and Decision Analysis* **2** 152–167.

KLEIN, N., KNEIB, T., LANG, S. and SOHN, A. (2015). Bayesian structured additive distributional regression with an application to regional income inequality in Germany. *Ann. Appl. Stat.* **9** 1024–1052. MR3371346

LIN, Y. and JEON, Y. (2006). Random forests and adaptive nearest neighbors. *J. Amer. Statist. Assoc.* **101** 578–590. MR2256176

LONG, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. Sage Publications, Thousand Oaks, CA.

MEINSHAUSEN, N. (2006). Quantile regression forests. *J. Mach. Learn. Res.* **7** 983–999. MR2274394

MESSNER, J. W., MAYR, G. J. and ZEILEIS, A. (2016). Heteroscedastic censored and truncated regression with crch. *The R Journal* **8** (1) 173–181.

MESSNER, J. W., MAYR, G. J. and ZEILEIS, A. (2017). Non-homogeneous boosting for predictor selection in ensemble post-processing. *Mon. Weather Rev.* **145** 137–147.

NELDER, J. A. and WEDDERBURN, R. W. M. (1972). Generalized linear models. *J. R. Stat. Soc. Ser. A* **135** 370–384.

RASP, S. and LERCH, S. (2018). Neural networks for post-processing ensemble weather forecasts. *Mon. Weather Rev.* **146** 3885–3900.

RIGBY, R. A. and STASINOPOULOS, D. M. (2005a). Generalized additive models for location, scale and shape. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **54** 507–554. MR2137253

ROBINSON, N., REGETZ, J. and GURALNICK, R. P. (2014). EarthEnv-DEM90: A nearly-global, void-free, multi-scale smoothed, 90m digital elevation model from fused ASTER and SRTM data. *ISPRS J. Photogramm. Remote Sens.* **87** 57–67.

SCHEUERER, M. and HAMILL, T. M. (2015). Statistical post-processing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Mon. Weather Rev.* **143** 4578–4596.

SCHLOSSER, L., HOTHORN, T., STAUFFER, R. and ZEILEIS, A. (2019a). Different response distributions. Supplement A to "Distributional regression forests for probabilistic precipitation forecasting in complex terrain." DOI:10.1214/19-AOAS1247SUPPA.

SCHLOSSER, L., HOTHORN, T., STAUFFER, R. and ZEILEIS, A. (2019b). Stationwise evaluation. Supplement B to "Distributional regression forests for probabilistic precipitation forecasting in complex terrain." DOI:10.1214/19-AOAS1247SUPPB.

STASINOPOULOS, D. M. and RIGBY, R. A. (2007). Generalized additive models for location scale and shape (GAMLSS) in R. *J. Stat. Softw.* **23** (7) 1–46.

STAUFFER, R., UMLAUF, N., MESSNER, J. W., MAYR, G. J. and ZEILEIS, A. (2017a). Ensemble post-processing of daily precipitation sums over complex terrain using censored high-resolution standardized anomalies. *Mon. Weather Rev.* **45** 955–969.

STAUFFER, R., MAYR, G. J., MESSNER, J. W., UMLAUF, N. and ZEILEIS, A. (2017b). Spatio-temporal precipitation climatology over complex terrain using a censored additive regression model. *Int. J. Climatol.* **37** 3264–3275.

STIDD, C. K. (1973). Estimating the precipitation climate. *Water Resour. Res.* **9** 1235–1241.

STRASSER, H. and WEBER, CH. (1999). The asymptotic theory of permutation statistics. *Math. Methods Statist.* **8** 220–250. Johann Pfanzagl—On the occasion of his 70th birthday. MR1722622

UGARTE, M. D., ADIN, A. and GOICOA, T. (2017b). One-dimensional, two-dimensional, and three dimensional B-splines to specify space-time interactions in Bayesian disease mapping: Model fitting and model identifiability. *Spat. Stat.* **22** 451–468. MR3732861

WOOD, S. N., SCHEIPL, F. and FARAWAY, J. J. (2013). Straightforward intermediate rank tensor product smoothing in mixed models. *Stat. Comput.* **23** 341–360. MR3041440

ZEILEIS, A. and HORNIK, K. (2007). Generalized M-fluctuation tests for parameter instability. *Stat. Neerl.* **61** 488–508. MR2351461

ZEILEIS, A., HOTHORN, T. and HORNIK, K. (2008). Model-based recursive partitioning. *J. Comput. Graph. Statist.* **17** 492–514. MR2439970

L. SCHLOSSER
R. STAUFFER
A. ZEILEIS
DEPARTMENT OF STATISTICS
FACULTY OF ECONOMICS AND STATISTICS
UNIVERSITÄT INNSBRUCK
UNIVERSITÄTSSTR. 15
6020 INNSBRUCK
AUSTRIA
E-MAIL: Lisa.Schlosser@uibk.ac.at
Reto.Stauffer@uibk.ac.at
Achim.Zeileis@R-project.org
URL: https://www.uibk.ac.at/statistics/personal/schlosser-lisa/
https://retostauffer.org/
https://eeecon.uibk.ac.at/~zeileis/

T. HOTHORN
INSTITUT FÜR EPIDEMIOLOGIE,
BIOSTATISTIK UND PRÄVENTION
UNIVERSITÄT ZÜRICH
HIRSCHENGRABEN 84
CH-8001 ZÜRICH
SWITZERLAND
E-MAIL: Torsten.Hothorn@R-project.org

Article VI

Gebetsberger M., Stauffer R., Mayr G.J., and Zeileis, A. (2019). *Skewed Logistic Distribution for Statistical Temperature Post-Processing in Mountainous Areas*. Advances in Statistical Climatoloy, Meteorology and Oceanography, 5, 87–100, doi:10.5194/ASCMO-5-87-2019.

Recent peer-reviewed journal on the intersection of atomspheric science and statistics (published by Copernicus), not yet listed in JCR.

Contribution (CRT): *Conceptualization / data curation / formal analysis / investigation / software / validation / writing, original draft*.

Adv. Stat. Clim. Meteorol. Oceanogr., 5, 87–100, 2019
https://doi.org/10.5194/ascmo-5-87-2019

ASCMO
Open Access

Skewed logistic distribution for statistical temperature post-processing in mountainous areas

Manuel Gebetsberger[1,2,3], **Reto Stauffer**[4], **Georg J. Mayr**[1], **and Achim Zeileis**[4]

[1]Institute of Atmospheric and Cryospheric Sciences, University of Innsbruck, Innsbruck, Austria
[2]LuftBlick, Innsbruck, Austria
[3]Division for Biomedical Physics, Medical University of Innsbruck, Innsbruck, Austria
[4]Department of Statistics, University of Innsbruck, Innsbruck, Austria

Correspondence: Manuel Gebetsberger (manuel.gebetsberger@gmail.com)

Received: 10 May 2018 – Revised: 4 April 2019 – Accepted: 15 May 2019 – Published: 18 June 2019

Abstract. Nonhomogeneous post-processing is often used to improve the predictive performance of probabilistic ensemble forecasts. A common quantity used to develop, test, and demonstrate new methods is the near-surface air temperature, which is frequently assumed to follow a Gaussian response distribution. However, Gaussian regression models with only a few covariates are often not able to account for site-specific local features leading to uncalibrated forecasts and skewed residuals. This residual skewness remains even if many covariates are incorporated. Therefore, a simple refinement of the classical nonhomogeneous Gaussian regression model is proposed to overcome this problem by assuming a skewed response distribution to account for possible skewness. This study shows a comprehensive analysis of the performance of nonhomogeneous post-processing for the 2 m temperature for three different site types, comparing Gaussian, logistic, and skewed logistic response distributions. The logistic and skewed logistic distributions show satisfying results, in particular for sharpness, but also in terms of the calibration of the probabilistic predictions.

1 Introduction

Probabilistic weather forecasts have become state-of-the-art in recent years (Gneiting and Katzfuss, 2014). As such, they are important for addressing the chaotic nature of the atmosphere and expressing the uncertainty of a specific forecast (Lorenz, 1963). The expected uncertainty is typically provided by an ensemble prediction system (EPS; Leith, 1974) where multiple forecasts are produced by a numerical weather prediction (NWP) model with slightly perturbed initial conditions, model physics, and parameterizations. However, it was found that these forecasts often show systematic errors in both the expectation and the uncertainty due to required simplified physical equations, insufficient resolution, and unresolved processes (Bauer et al., 2015).

Statistical post-processing techniques (Gneiting and Katzfuss, 2014), such as Gaussian ensemble dressing (GED; Roulston and Smith, 2003), nonhomogeneous Gaussian regression (NGR or EMOS; Gneiting et al., 2005), a nonhomogeneous mixture model approach with similarities to Bayesian model averaging (BMA; Raftery et al., 2005), or logistic regression (Wilks, 2009; Messner et al., 2014), are one possibility to correct for these errors. These methods have been extensively tested for air temperature forecasts and other quantities, with NGR (with various extensions) representing one of the most popular approaches.

The two most important properties of probabilistic forecasts are sharpness and calibration (Gneiting et al., 2007) which have to be considered jointly. Accurate forecasts should be as sharp as possible but not overconfident, as this would result in a loss of calibration. Previous studies show that extensions of the classical NGR method (Schefzik et al., 2013; Scheuerer and Büermann, 2014; Möller and Groß, 2016; Dabernig et al., 2017) and other temperature post-processing methods (Hagedorn et al., 2008; Verkade et al., 2013; Feldmann et al., 2015; Wilks, 2017) are able to improve the predictive performance of the classical NGR with respect to specific predictive performance measures such as sharpness and calibration.

However, in recent publications, the probability transform histograms (PIT; Dawid, 1984) presented often do not show the desired perfectly uniform distribution to confirm calibration (cf., Scheuerer and Büermann, 2014, Fig. 5c,g; Möller and Groß, 2016, Fig. 4c; or Messner et al., 2017, Fig. 7). More specifically, the histograms indicate skewness in the residual distribution. As a marginal Gaussian model without covariates can already exhibit skewness for temperature data (Toth and Szentimrey, 1990; Warwick and Curran, 1993; Harmel et al., 2002), skewness is supposed to vanish if covariates are incorporated. Nevertheless, the residual distribution is still found to be skewed even after adjustment using covariates (Messner et al., 2017). As covariates are based on the output of NWP models, a remaining skewness is likely to originate in small-scale or local atmospheric processes that are insufficiently or not at all resolved by the NWP models. Locations in regions where topography is only coarsely resolved in the model are an example of this. As a result, many thermally induced slope and valley wind systems as well as subsidence/lifting zones (Steinacker, 1984; Whiteman, 1990; Zängl, 2004) will be absent, which may cause residual skewness in the post-processed forecasts.

So far, most studies assume a Gaussian response distribution for their temperature post-processing methods (Gneiting et al., 2005; Hagedorn et al., 2008; Verkade et al., 2013; Scheuerer and Büermann, 2014; Möller and Groß, 2016; Gebetsberger et al., 2018; Dabernig et al., 2017). As the Gaussian distribution is symmetric, it is not able to account for possible skewness by itself. Hence, this article proposes an extension of the nonhomogeneous Gaussian regression framework (Gneiting et al., 2005) using a skewed rather than a symmetric response distribution in order to obtain sharp and calibrated probabilistic temperature forecasts. To examine the need for asymmetry, probabilistic temperature forecasts are presented for a set of stations with different characteristics including sites in the European Alps and plain areas across central Europe. Moreover, the current study uses a long-term data set for training the statistical models, and compares the results to the widely used sliding training period approach where a fixed number of past training days is used (Gneiting et al., 2005; Scheuerer and Büermann, 2014; Feldmann et al., 2015; Möller and Groß, 2016).

2 Methods and data

Section 2.1 briefly describes the regression framework followed by the response distributions as used in this study (Sect. 2.2). The data and statistical model specifications are introduced in Sect. 2.3, and the verification methodology to access the predictive performance is introduced in Sect. 2.4.

2.1 Nonhomogeneous regression framework

The nonhomogeneous Gaussian regression (NGR) framework as proposed by Gneiting et al. (2005) is a special case of a distributional regression model (Klein et al., 2015) and can be expressed in its general form as

$$y \sim \mathcal{D}(h_1(\theta_1) = \eta_1, \ldots, h_K(\theta_K) = \eta_K). \quad (1)$$

A response variable y is assumed to follow some probability distribution $\mathcal{D}$ with distribution parameters θ_k, $k = 1, \ldots, K$. Each parameter is linked to an additive predictor η_k using a monotone link function h_k. In this article we use the identity-link $h_k(\eta_k) = \eta_k$ for the location parameter and a log-link for scale and shape parameters to ensure positivity during optimization, as proposed in Gebetsberger et al. (2017). Each linear predictor can be expressed by a set of additive predictors which have the following form:

$$\eta_k = \eta_k(\mathbf{x}_p, \boldsymbol{\beta}_k) = f_{1k}(\boldsymbol{x}_1, \beta_{1k}) + \ldots + f_{Pk}(\boldsymbol{x}_P, \beta_{Pk}), \quad (2)$$

including various (possibly nonlinear) functions f_{pk}, $p = 1, \ldots, P$. Hence, $\mathbf{x}_p$ defines a matrix of covariates used, and $\boldsymbol{\beta}_{pk}$ is the vector of the regression coefficients to be estimated.

Classical NGR (Gneiting et al., 2005) assumes the Gaussian response distribution, which is described by the two parameters for location and scale. In ensemble post-processing applications, it is common to use the ensemble covariate which describes the observed variable of interest, e.g., ensemble temperature is used for temperature observations. The term nonhomogeneous relates to the residual variance, which, in contrast to linear (homogenous) regression, varies depending on the covariate value used for the Gaussian scale parameter (Wilks, 2011). The optimization of the regression coefficients is originally carried out by minimizing the continuous ranked probability score (CRPS; Hersbach, 2000), although it can also be estimated by maximum likelihood estimation (ML; Aldrich, 1997). Both approaches are compared in Gebetsberger et al. (2018), where it is shown that CRPS optimization obtains sharper, but not necessarily better, calibrated probabilistic predictions than ML estimation.

2.2 Response distributions

This study compares three different distributions for temperature post-processing: (i) the frequently-used Gaussian distribution, (ii) the symmetric logistic distribution, and (iii) the generalized logistic distribution type I (Fig. 1). The logistic distribution is used to assess the impact of having slightly heavier tails (Gebetsberger et al., 2018). The generalized logistic distribution type I is of particular interest as it allows one to account for possible skewness in the data. For simplicity, it will be referred as the skewed logistic distribution in the following.

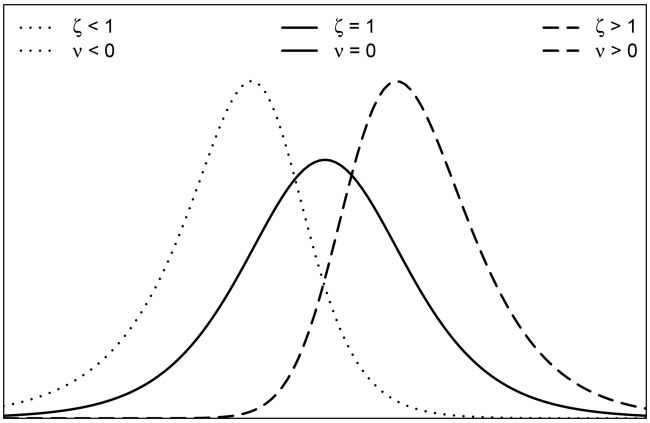

Figure 1. Density function of the skewed logistic distribution, illustrating the third moment (ν, skewness) depending on the chosen shape parameter ζ.

The skewed logistic distribution has the cumulative distribution function (CDF):

$$\mathrm{CDF}(x) = \frac{1}{\left(1+\exp(-\frac{x-\mu}{\sigma})\right)^{\zeta}}, \tag{3}$$

with location parameter μ, scale parameter σ, and shape parameter ζ. The first derivation of Eq. (3) leads to the probability density function (PDF):

$$\mathrm{PDF}(x) = \frac{\zeta \cdot \exp\left(-\frac{x-\mu}{\sigma}\right)}{\sigma \cdot \left(1+\exp\left(-\frac{x-\mu}{\sigma}\right)\right)^{2\zeta}}. \tag{4}$$

The additional shape parameter ζ is responsible for the skewness. Figure 1 shows the PDF for three different shape parameter values of ζ and corresponding skewness ν. ζ is positive where values below 1 create negative skewness (heavier left tail, $\nu < 0$), whereas values above 1 produce positive skewness (heavier right tail, $\nu > 0$). For $\zeta \equiv 1$ the skewed logistic distribution describes the symmetric logistic distribution.

As an example, values for $\zeta = \{0.50, 1, 3.82\}$ produce a skewness of $\nu = \{-0.85, 0, 0.85\}$ as illustrated in Fig. 1. Details regarding the skewness calculation can be found in Appendix A.

2.3 Data and statistical models

2.3.1 Data

Results are presented at 27 different sites in central Europe (Fig. 2) for forecasts +12 to +96 h at 6-hourly intervals. The sites were selected to investigate the influence of different topographical environments. Therefore, the stations are subjectively clustered into three distinct groups representing Alpine sites located in inner-Alpine regions (12), foreland sites in the peripheral area close to the Alps (6), and plain sites in topographically flat areas (9). Nevertheless, statistical models

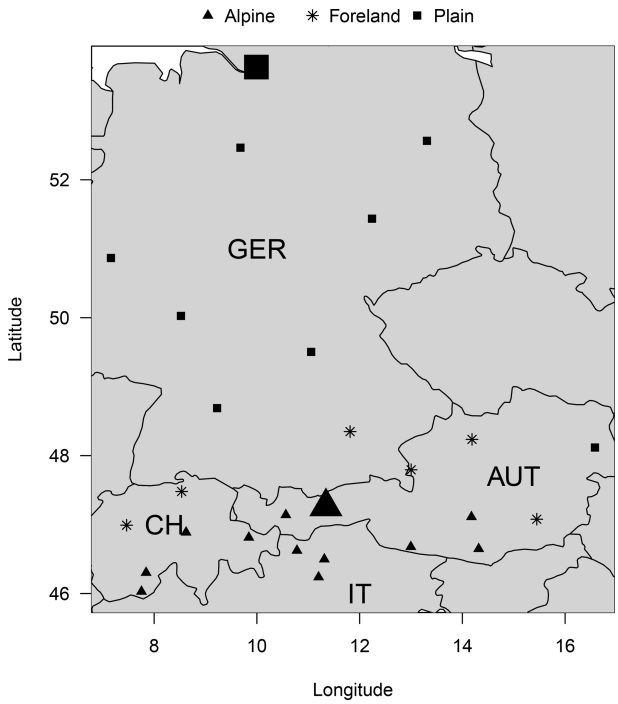

Figure 2. Study area and selected stations in Germany (GER), Switzerland (CH), Italy (IT), and Austria (AUT). The markers indicate stations classified as Alpine (triangle), foreland (star), and plain (square). Large symbols represent stations that are discussed in detail in this article: Innsbruck, Austria (large triangle), and Hamburg, Germany (large square).

described in the next subsection are estimated individually for each station and lead time, as each location and time of the day has its own site-specific characteristics.

Temperature observations are provided by automatic weather stations (10 min mean values). As input, 2 m temperature forecasts of the 50 + 1 member EPS of the European Centre for Medium-Range Weather Forecasts (ECMWF) are used. For this study only EPS forecasts initialized at 00:00 UTC are considered. The data set covers the time period from 1 January 2012 to 31 December 2015 resulting in 4 years of data that yield a sample size of approximately 1400 for each individual station and forecast lead time. The temperature covariate of the raw ECMWF ensemble is bilinearly interpolated to the individual sites.

In this article, detailed case studies will be shown for Innsbruck, Austria (Alpine site), and Hamburg, Germany (plain site; cf., Fig. 2), which differ – particularly with respect to their topographical environments. While the Alpine site is located in a narrow Alpine valley surrounded by high mountainous exceeding an altitude of 2500 m, the plain site is characterized by its proximity to the sea (100 km), its few hills, and an altitude below 160 m. Due to the necessary simplifications in the NWP model, the topography is missing large parts of the topographical structures, especially for the Alpine site (Stauffer et al., 2017, Figs. 1 and 3).

2.3.2 Statistical models

Similar to previous works (cf., Scheuerer and Büermann, 2014; Feldmann et al., 2015; Möller and Groß, 2016; Dabernig et al., 2017), we only utilize the ensemble mean ($\overline{\text{ens}}$) and ensemble standard deviation (SD_{ens}) of the 2 m temperature forecasts from the ECMWF EPS in this study. In the following model specification, the ensemble mean is used for the linear predictor of the location parameter μ, whereas the ensemble standard deviation is used for the linear predictor of the corresponding scale parameter σ.

While the Gaussian and the logistic distribution have only two parameters (μ and σ), the skewed logistic distribution has an additional shape parameter ζ. To be able to capture seasonality, a smooth cyclic spline f depending on the day of the year (DOY) is used in the linear predictor for all distribution parameters. The seasonal splines allow the regression coefficients to vary over the year, if needed, while the cyclic constraint avoids discontinuities at the turn of the year. As there is no obvious candidate among all of the parameters provided by the EPS, the shape parameter ζ of the skewed logistic distribution is solely expressed by a smooth cyclic spline. This allows the model to account for possible skewness in the residuals between the observed and forecasted 2 m temperature. The model specification for the study presented can be summarized as follows:

$$y \sim \mathcal{D}(\mu, \sigma, \zeta), \tag{5}$$

$$\mu = f(\text{DOY}) + \beta_1 \cdot \overline{\text{ens}}, \tag{6}$$

$$\log(\sigma) = f(\text{DOY}) + \gamma_1 \cdot \log(\text{SD}_{\text{ens}}), \tag{7}$$

$$\log(\zeta) = f(\text{DOY}), \tag{8}$$

for which the additional parameter ζ is solely used for models utilizing the skewed logistic response distribution. The optimization of the regression coefficients for all parameters is performed employing likelihood based gradient boosting (R package "bamlss"; Umlauf et al., 2018). In this context gradient boosting is not used for variable selection, but to obtain regularized estimates for the regression coefficients. This is done by performing an additional 10-fold cross validation on the training data set to find the optimal stopping criterion based on the 10-fold out-of-sample root mean squared error. Table 1 shows a comprehensive overview of all of the models and the covariates used in the corresponding linear predictors.

2.4 Verification methodology

Different scores are used to assess the predictive performance of the models tested. The overall performance is evaluated by the logarithmic score (LS; Wilks, 2011) and the continuous ranked probability score (CRPS; Hersbach, 2000). The LS evaluates a forecast distribution by taking the logarithmic probability density value at the observed value, whereas the CRPS accounts for the whole forecast distribution.

Table 1. Covariates used in the linear predictors of the distributional parameters μ, σ, and ζ for all response distributions. $\overline{\text{ens}}$ and SD_{ens} represent the ensemble mean and the standard deviations of the ensemble 2 m temperature, respectively; $f(\text{DOY})$ represents the smooth cyclic seasonal effect.

Name/ distribution	μ	$\log(\sigma)$	$\log(\zeta)$
Gaussian	$f(\text{DOY})$, $\overline{\text{ens}}$	$f(\text{DOY})$, SD_{ens}	–
Logistic	$f(\text{DOY})$, $\overline{\text{ens}}$	$f(\text{DOY})$, SD_{ens}	–
Skewed logistic	$f(\text{DOY})$, $\overline{\text{ens}}$	$f(\text{DOY})$, SD_{ens}	$f(\text{DOY})$

Of particular interest for this study is the performance of the post-processing models in terms of sharpness and calibration (Gneiting et al., 2007). The sharpness of the probabilistic forecasts is verified using the average prediction interval width (PIW). Results for three different intervals are shown in this article: 50 %, 80 %, and 95 %. For example, the 80 % PIW describes the range between the 10th percentile and the 90th percentile of the probabilistic forecast. The smaller the PIW, the sharper the forecasts.

Calibration is visually evaluated using probability integral transform (PIT) histograms (Gneiting et al., 2007), which evaluate the forecasted cumulative distribution functions equivalent to the rank histogram (Anderson, 1996; Talagrand et al., 1997; Hamill and Colucci, 1998). In addition, the reliability index (RI; Delle Monache et al., 2006) and prediction interval coverage (PIC) are shown. The RI allows one to analyze an aggregated measure over a large number of individual PIT histograms. RIs are defined as $\sum_{i=1}^{I} |\kappa_i - \frac{1}{I}|$, where I defines the number of individual bins in a PIT histogram and κ_i defines the observed relative frequency in each bin. In this study we use a binning of 5 %. The RI describes the sum of the absolute deviation from each bin in a specific PIT histogram from perfect calibration. Thus, perfectly calibrated forecasts would show an RI of zero. PICs show the calibration for a specific interval. As for the PIW, PICs are shown for the 50 %, 80 %, and 95 % interval in addition to theoretical PICs of 50 %, 80 %, and 95 %. The closer the empirical PIC is to the theoretical PIC, the better the calibration.

3 Results and discussion

This section presents a detailed analysis of the different statistical models. Section 3.1 shows a detailed analysis of the long-term training window approach (see Sect. 2.3) for an Alpine valley site. These results are compared to the results for a plain site in Sect. 3.2. Section 3.3 shows a comprehensive analysis of the predictive performance of the proposed method for the three different groups of stations (Alpine, foreland, and plain sites), whereas Sect. 3.4 compares the proposed long-term training data approach against the frequently used sliding window approach.

All results presented in Sect. 3.1–3.2 are out-of-sample results using 4-fold block-wise cross-validation. For each model, station, and lead time, four individual regression models have been estimated using 3 years of data while one full year (2012, 2013, 2014, or 2015) is used as test data set. The comparison in Sect. 3.4 is based on out-of-sample results for the year 2015 if not stated otherwise. Sliding window models are estimated by minimum CRPS and maximum likelihood estimation as in Gebetsberger et al. (2018).

3.1 Alpine case study

Raw ensemble forecasts for Alpine sites cannot be directly used because the topography is not well resolved. Therefore, raw ensemble forecasts are typically characterized by small 80 % prediction interval widths (PIWs) around 3 °C and large CRPS values around 4 (Stauffer et al., 2018). The large CRPS values are mainly driven by a systematic bias because of the difference between the real and model topography of the ECMWF. Additionally, the small PIW of the raw ensemble leads to underdispersive probabilistic predictions (Gebetsberger et al., 2018).

To show the performance of the proposed approach, the analysis for one selected site with a distinct Alpine character is shown (large triangle, Fig. 2). The left column of Fig. 3 presents the verification for this Alpine site. Figure 3 (top down) shows LS, CRPS, 80 % PIW, and RI for all forecast lead times. A dominant diurnal cycle for LS, CRPS, and the 80 % PIW can be seen for all three models, with the smallest (best) scores obtained during nighttime (00:00 and 06:00 UTC) and largest during daytime (12:00 and 18:00 UTC). The increased PIW during nighttime in combination with low RIs show that forecasts at night are sharper than during the day, although both are well calibrated. Overall, only a small decrease in the forecast performance can be identified with increasing lead time which implies comparable skill between the first and fourth forecast day.

When comparing the logistic model with the benchmark Gaussian model, the logistic model shows small improvements in LS, especially during nighttime. Similar behavior can be seen for the sharpness (80 % PIW) where the strongest improvements can be achieved during nighttime, but with an overall improvement for all lead times. Furthermore, the logistic model is able to remove large parts of the existing diurnal pattern in terms of calibration, showing a more homogeneous RI for all lead times time compared with the Gaussian model. The proposed skewed logistic model shows similar performance in all verification measures compared to the logistic model, with the largest improvements in sharpness during nighttime.

Figure 4 shows PIT histograms for the 2 d ahead forecasts. To increase readability, only the Gaussian and skewed logistic models are shown. PITs are shown for 06:00, 12:00, 18:00, and 00:00 UTC to assess the characteristics for different times of the day. Top down PITs for the summer season, the winter season, and the full year are shown to highlight seasonal differences in calibration. Forecasts for day one, three, and four show a very similar picture (not shown).

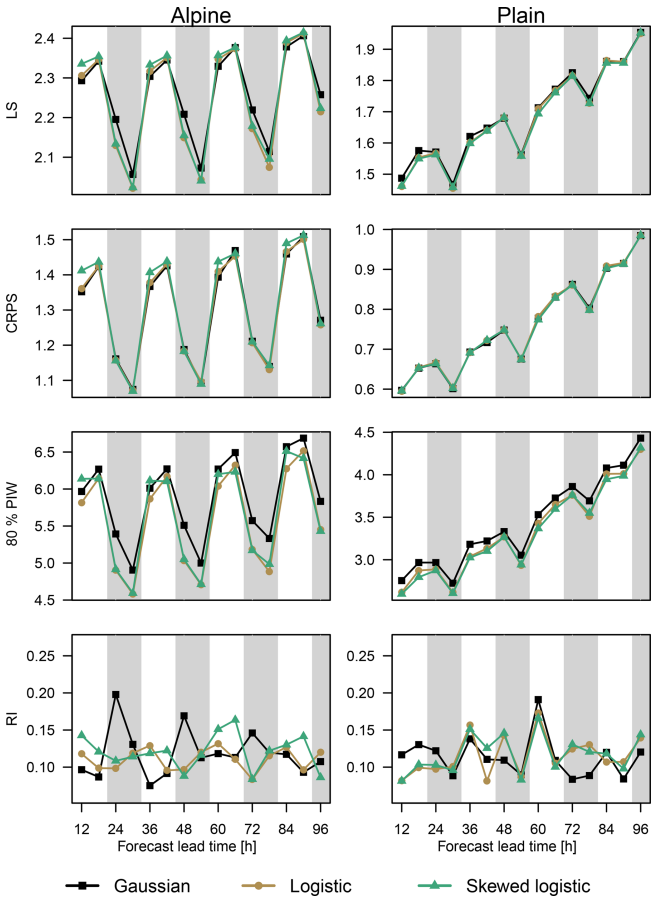

Figure 3. Performance measures at the selected Alpine site (left) and plain site (right) for all three models (Gaussian: squares; logistic: circles; skewed logistic: triangles). From the top down, the LS, CRPS, 80 % PIW, and RI are shown, and are evaluated for all forecasts +12 to +96 h ahead. Nighttime forecasts (00:00 and 06:00 UTC) are highlighted using vertical gray bars. Please note that the displayed range on the ordinate differs between the left and right column, except for RI.

Both, the Gaussian and logistic model, already show an almost uniform distribution, although for one particular hour of the day special features can be identified. The convex shape of the Gaussian model for the all year period at +48 h (bottom right) indicates overdispersion (peak at bin 0.5), while the asymmetry also indicates residual skewness (peak at bin 0.95). This is likely caused by the not yet resolved topography in the NWP. The overdispersion is more visible in the summer season for +48 h (Fig. 4, top right), where a peak can be seen at around 0.5, and two minima occur at 0.05 and 0.9, respectively. The skewed logistic distribution is able

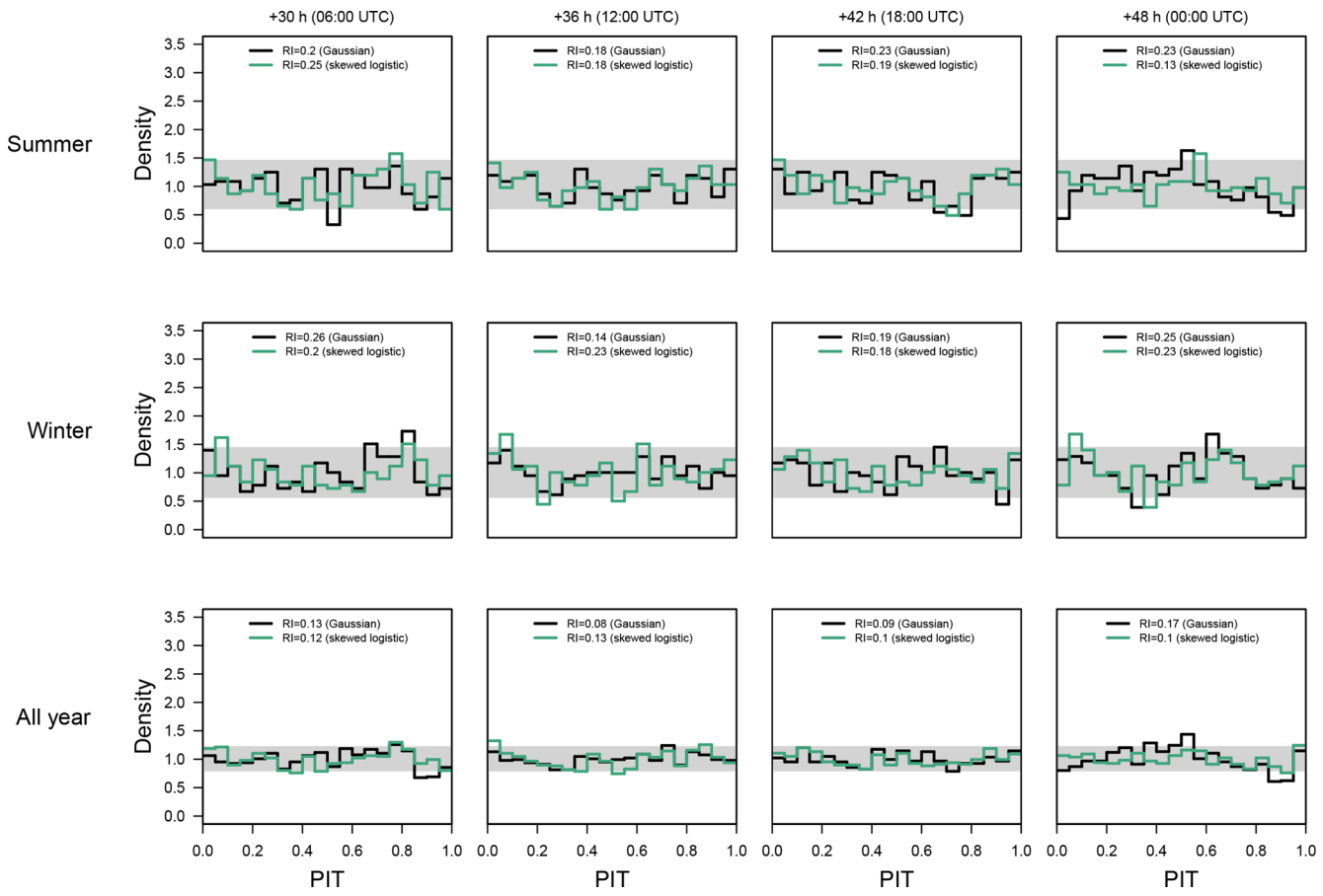

Figure 4. PIT histograms at the Alpine site for the Gaussian (black/dark line) and skewed logistic (green/bright line) models for the 2 d ahead forecasts (left to right: 06:00, 12:00, 18:00, and 00:00 UTC) corresponding to forecasts +30, +36, +42, and +48 h ahead. Top down, the PIT histograms are shown for summer only (June/July/August), winter only (December/January/February), and for the whole year. The gray horizontal bar shows the point-wise 95 % confidence interval around 1 which indicates perfect calibration.

to produce a more uniform PIT which is also quantified by smaller RI values.

Fig. 5a shows a joint time series of the empirical skewness for the skewed logistic models for all +36 h forecasts (12:00 UTC) over the whole validation period. The estimated seasonal effect for skewness based on the 4-fold cross-validation is plotted against the left-out year, the year which has not been used when estimating the model. Thus, the effects for the four years (2012–2015) look slightly different as they are based on four different models. However, the overall pattern across years is similar, which is an indication that this is a rather persistent characteristic given the data set used in this study. For all years the predictions are positively skewed during the summer season with values of around 0.6. On the contrary, strong negative skewness with values of −0.8 can be seen during the winter season. The consideration of this seasonally dependent skewness yields an overall better performance compared with the Gaussian model.

3.2 Alpine vs. plain site

To see the benefits of a nonsymmetric response distribution in a different environment, the same study is shown for a selected plain site (large square Fig. 2; right column Fig. 3).

Similar to the Alpine site, a pronounced diurnal cycle is visible for all models in terms of LS and CRPS (Fig. 3) with better scores for nighttime. In contrast to the Alpine site, a clear decrease in the forecast performance with increasing lead time can be seen; however, the two heavy-tailed models (logistic and skewed logistic) are still able to improve sharpness (80 % PIW) and calibration (RI) for particular lead times. The estimated skewness is also smaller than for the Alpine site, as shown in Fig. 5. Additionally, the change in sign of the skewness between summer and winter is almost absent. Skewness is still present, but the amplitude is strongly decreased compared with the results for the Alpine site with values of close to zero (symmetric). Even if the improvements over the symmetric logistic models are only minor, the additional skewness still yields slightly better results, especially for short lead times.

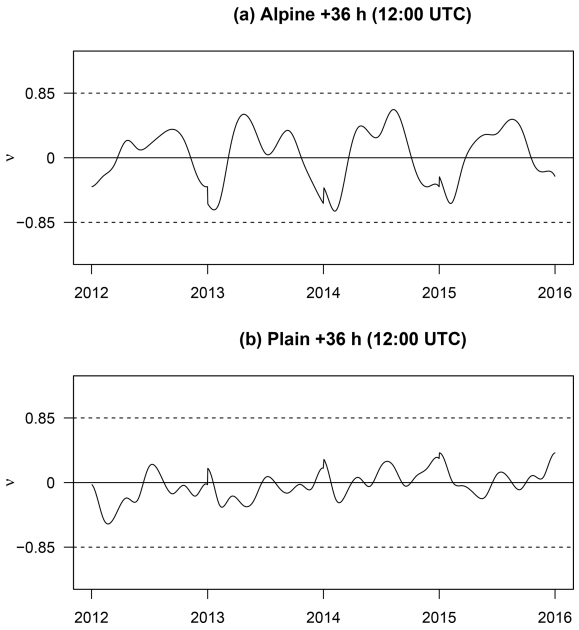

Figure 5. Joint time series of the empirical skewness ν of the forecasted skewed logistic distribution for a lead time of +36 h for the Alpine station **(a)** and plain site **(b)** for all four cross-validation blocks. The years on the abscissa correspond to the left-out year of the cross-validation. Symmetric forecasts (no additional skewness required) would show a value of zero. The two dashed lines at $\nu = [-0.85, 0.85]$ are somewhat arbitrary and are shown to facilitate orientation and to match the examples in Fig. 1.

In comparison with the Alpine site, the plain site shows an overall better forecast performance for all measures except for RI where both stations show similar scores indicating that both stations are, on average, well calibrated. Moreover, almost all scores (LS, CRPS, and PIW) are smaller than for the Alpine site even for the longest lead time. This is mainly due to the overall better performance of the NWP for regions with no or few topographical features. In such situations the overall performance of the NWP is already adequate and the EPS provides covariates containing more information. Thus, the benefit of the statistical post-processing is much smaller compared with sites in complex terrain. In this example the Gaussian assumption seems to be an appropriate choice, and the improvements of the logistic or skewed logistic distribution are only minor.

3.3 Comparison for all sites

Figure 6 shows averaged scores for LS, CRPS, the mean 80 % PIW, and RI for the three different groups of stations including all 27 sites used in this study (cf. Fig. 2). Each box and whiskers contains the mean score for the individual stations and all 15 lead times. This yields 12×15 values for group "Alpine", 6×15 for group "foreland", and 9×15 for group "plain". In addition, the numeric values of all medians are provided in Table 2 along with median values for two alternative PIWs (50 % and 95 %) and the prediction interval coverage (PIC) for the same three intervals. The validation shows increasing forecast performance with decreasing topographical complexity (top down) independent of the statistical model.

Figure 7 shows the improvements using non-Gaussian distributions, compared with the Gaussian reference model: positive values indicate that the alternative model show an improvement over the Gaussian model. The model results using the symmetric/skewed logistic distribution show minor improvements in terms of LS but can clearly reduce the 80 % PIW without a loss in RI, except at Alpine sites where the logistic model shows a loss in RI (not as well calibrated). CRPS reports barely any difference between the different response distributions for all three groups. Large parts of the improvements can be attributed to the increased sharpness (PIW), which also yields a smaller LS overall without decreasing calibration in terms of RI.

3.4 Comparison to sliding training window

In the following, the long-term training approach presented using 3 years of training data (2012, 2013, and 2014) is compared to the widely used sliding window approach utilizing only the previous 30 or 60 d for training. The validation period chosen is 2015 in order to have at least 1 year of out-of-sample data. Skewed logistic models are not estimated for sliding windows. Due to the parametrization of the skewed logistic distribution and the relatively short training periods, reliable parameter estimates can no longer be ensured; therefore, only results for the Gaussian and logistic models are shown. The estimation of all sliding window models is based on the R package "crch" (Messner et al., 2016) using either minimum CRPS or maximum likelihood optimization (cf., Gebetsberger et al., 2018).

Figures 8 and 9 show overall scores and skill scores as in the previous subsection. The long-term approach using 3 years of training data shows the smallest LS and CRPS values for the entire validation period. Sharpness in terms of 80 % PIW is lowest for sliding window models. In particular, the sharpness is clearly lower than for long-term training models at Alpine sites. Moreover, the PIW is lower for short (30 d) than for longer (60 d) windows, especially for the 30 d sliding window models at the expense of calibration (RI). RI values report similar behavior, with 60 d sliding windows reporting smaller RI values than 30 d windows. As this verification is solely based on 1 year, there is large variation in the RI values, which is based on PIT histograms.

Therefore, Fig. 10 illustrates a representative PIT for the Alpine site, evaluated for the 60 d sliding window model using CRPS optimization, over the entire data period from March 2012 to the year 2015 (4 years minus 60 d). A distinct U-shape can be identified in the all year verification with peaks in the lowest and highest PIT bins. In particular, the

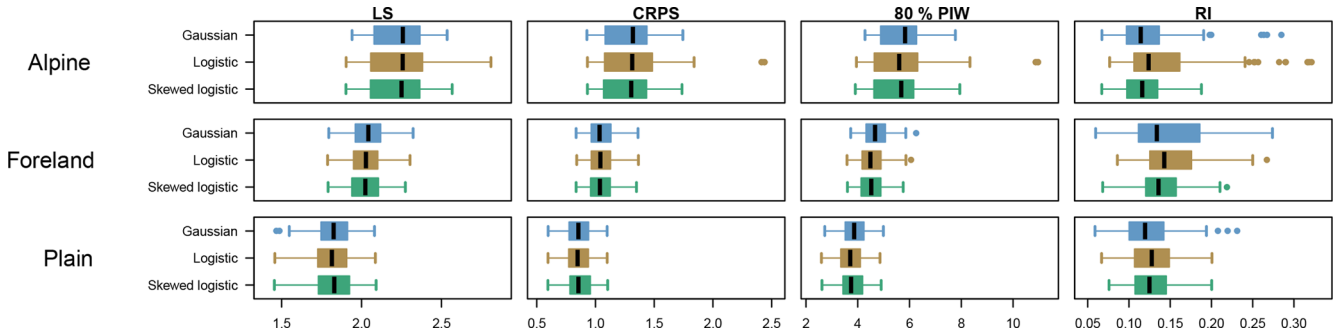

Figure 6. Performance measures in terms of LS, CRPS, 80 % PIW, and RI (left to right), clustered for Alpine, foreland, and plain sites (top to bottom). The box and whiskers are based on average scores for each station and lead time, with the boxes illustrating the interquartile range (0.25–0.75), the whiskers denoting ±1.5 times interquartile range, and the solid circles representing outliers.

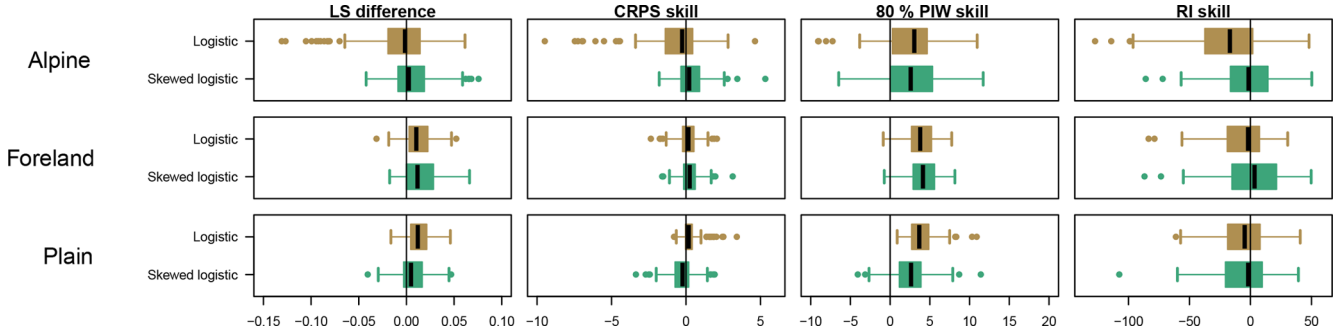

Figure 7. As in Fig. 6, but showing the improvement against the classical Gaussian model. Note that improvements are reported by positive values. Differences are shown for LS, whereas skill scores (in %) are shown for CRPS, the 80 % PIW, and the RI.

Table 2. Median of (left to right) the logarithmic score (LS), the continuous ranked probability score (CRPS), the reliability index (RI), and three prediction intervals (PIs) reporting the prediction interval width (PIW) and the prediction interval coverage (PIC) for Alpine, foreland, and plain sites (top to bottom), evaluated for each model type (Gaussian, logistic, and skewed logistic).

	Model	LS	CRPS	RI	PI 50 %		PI 80 %		PI 95 %	
					PIW	PIC	PIW	PIC	PIW	PIC
Alpine	Gaussian	2.26	1.32	0.11	3.07	50.24	5.83	79.40	8.91	93.92
	Logistic	2.26	1.31	0.12	2.80	46.55	5.60	78.22	9.34	95.02
	Skewed logistic	2.25	1.30	0.12	2.84	47.42	5.68	78.34	9.45	94.48
Foreland	Gaussian	2.04	1.04	0.13	2.46	50.87	4.67	80.28	7.14	94.06
	Logistic	2.03	1.04	0.14	2.24	47.08	4.49	79.03	7.48	94.93
	Skewed logistic	2.02	1.04	0.14	2.24	47.11	4.52	78.83	7.55	94.74
Plain	Gaussian	1.83	0.85	0.12	2.03	51.07	3.86	80.46	5.91	94.27
	Logistic	1.81	0.85	0.13	1.85	47.51	3.71	79.09	6.18	95.17
	Skewed logistic	1.83	0.85	0.12	1.87	47.89	3.74	79.23	6.25	95.30

sliding window approach shows a large peak in the lowest bin during summer, which also indicates residual skewness. Similar behavior is visible for winter periods, although it is less pronounced. The 60 d sliding window models using the maximum likelihood estimation decreases these peaks and yields more well-calibrated PITs (not shown) as they are less prone to being overconfident (Gebetsberger et al., 2018).

4 Summary and conclusion

Nonhomogeneous regression is a widely used statistical method for post-processing numerical ensemble forecasts. It was originally developed to improve probabilistic air temperature forecasts and assumes a Gaussian response distribution.

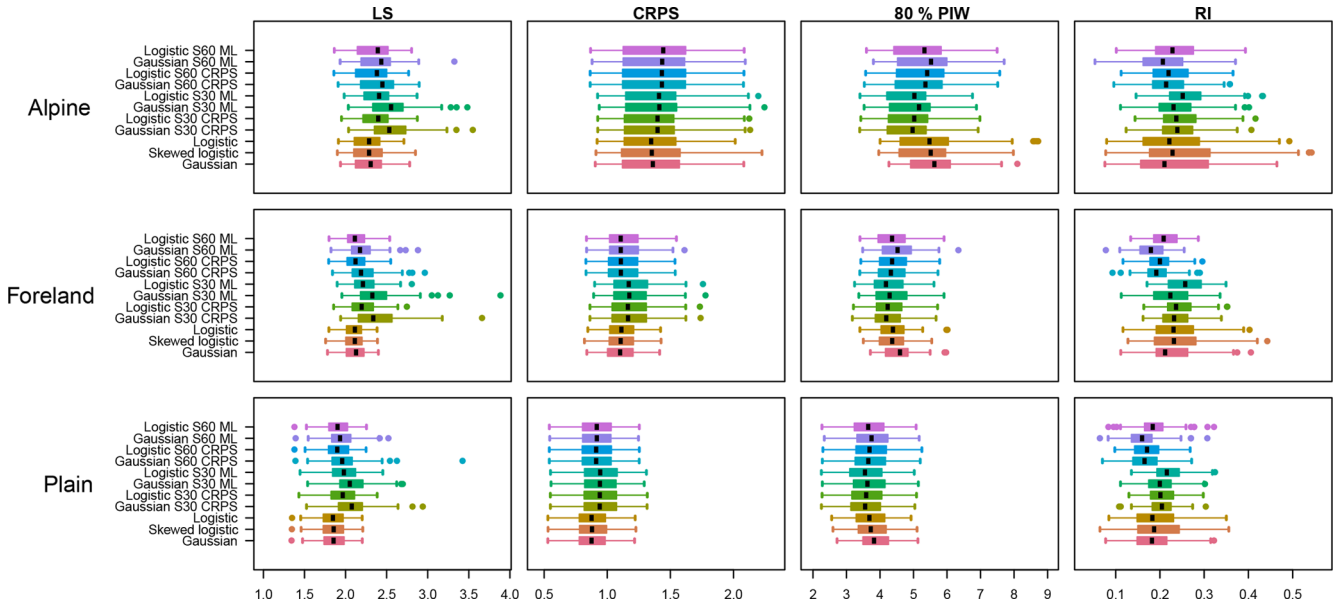

Figure 8. Performance measures in terms of LS, CRPS, 80 % PIW, and RI (left to right), clustered for Alpine, foreland, and plain sites (top to bottom), and were only evaluated on 2015 for out-of-sample comparison. The box and whiskers are based on average scores for each station and lead time, with boxes illustrating the interquartile range (0.25–0.75), whiskers displaying the ±1.5 times interquartile range, and solid circles representing outliers. Sliding training window models are labeled as S60 and S30 denoting a 60 or 30 d training period, respectively. Additionally, the optimization score used is labeled as CRPS or ML (continuous ranked probability score or maximum likelihood), respectively.

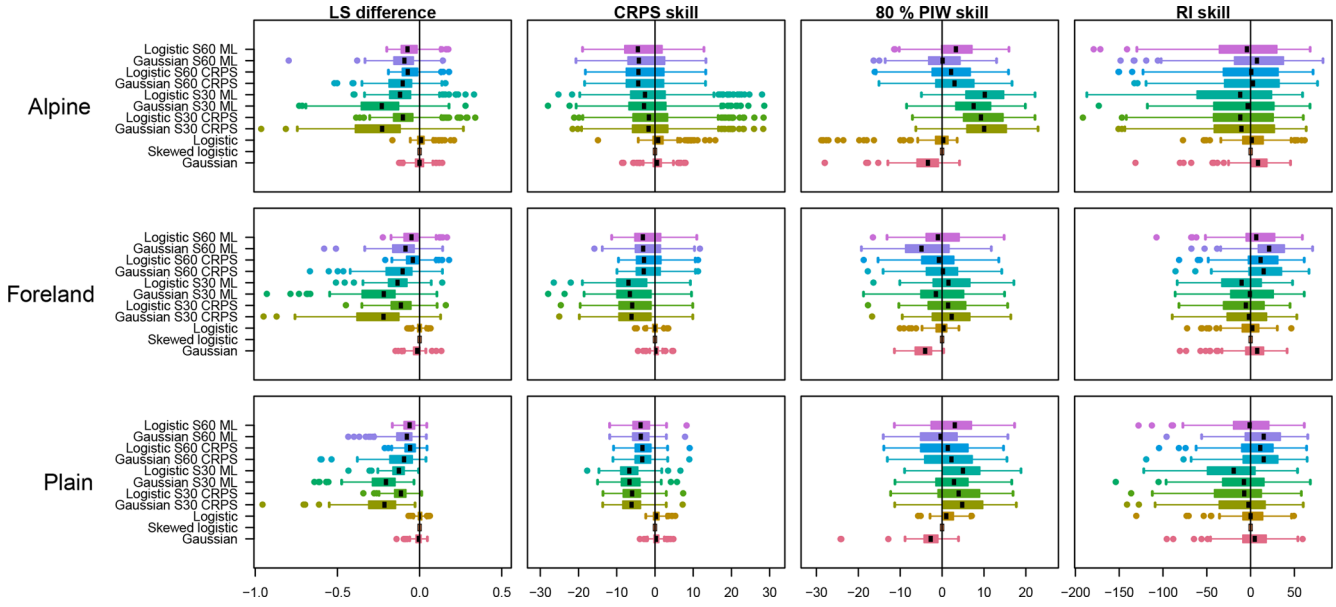

Figure 9. As in Fig. 8, but showing the improvements against the skewed logistic model. Note that improvements are reported by positive values. Differences are shown for LS, whereas skill scores (in %) are shown for CRPS, the 80 % PIW, and the RI.

However, several studies have shown that marginal temperature distributions can be skewed or nonsymmetric, respectively (Warwick and Curran, 1993; Harmel et al., 2002). This marginal skewness can result from topographically induced effects such as cold pools during winter or a strong valley bottom heating within narrow valleys on hot summer days. Thus, skewness is much stronger for locations surrounded by complex terrain than for sites in plain regions.

Moreover, skewness is supposed to decrease if additional covariates (e.g., individual ensemble members, seasonal ef-

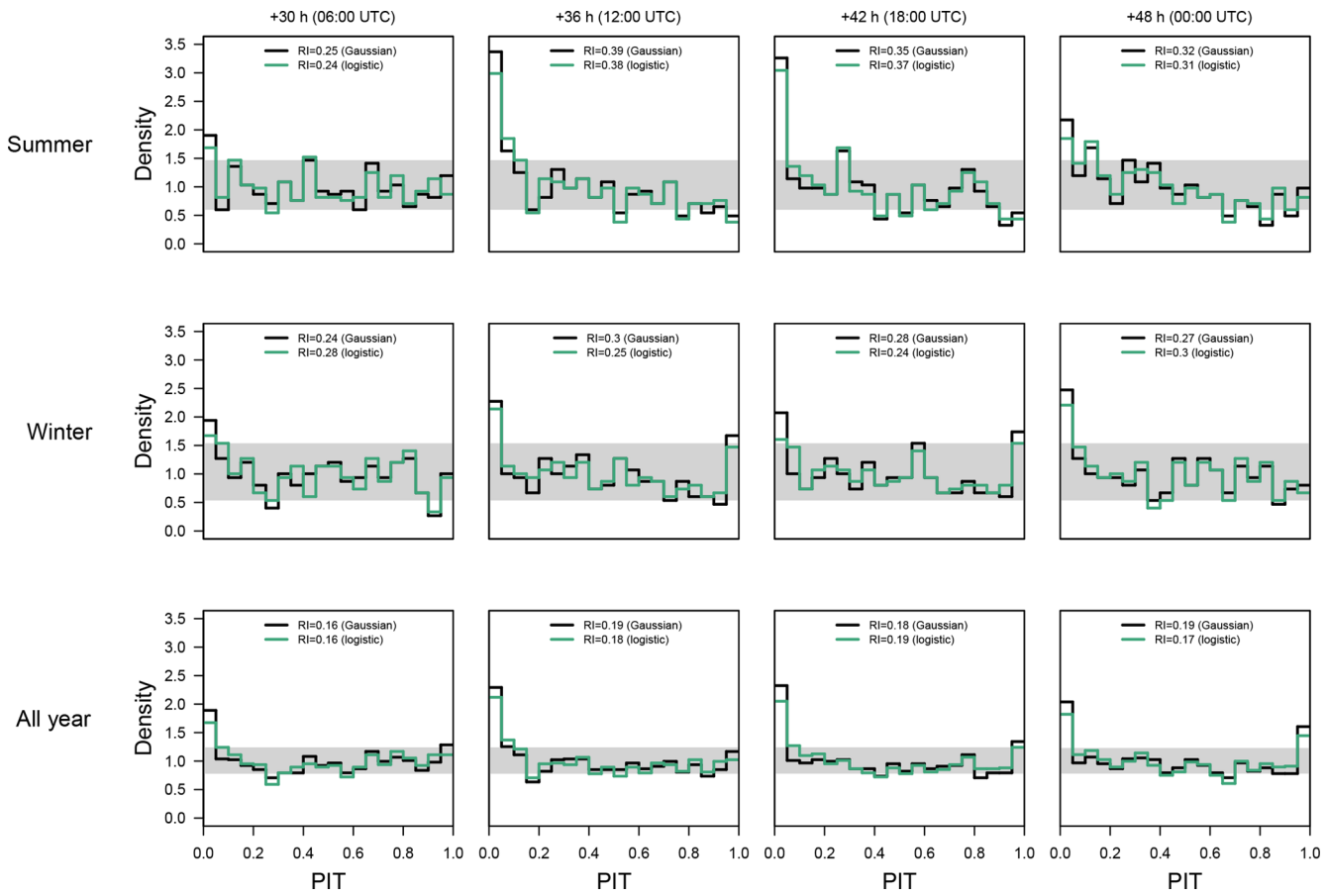

Figure 10. PIT histograms at the Alpine site for the Gaussian (black/dark line) and logistic (green/bright line) sliding 60 d models using CRPS optimization for the 2 d ahead forecasts (left to right: 06:00, 12:00, 18:00, and 00:00 UTC) corresponding to forecasts +30, +36, +42, and +48 h ahead. From the top down, the PIT histograms are shown for summer only (June/July/August), winter only (December/January/February), and for the whole year. The gray horizontal bar shows the point-wise 95 % confidence interval around 1 which indicates perfect calibration.

fect, and different ensemble forecast quantities) are included in the Gaussian model (see, e.g., Messner et al., 2017). However, the calibration of the results presented in this article indicate that residual skewness remains, even when including more variables than just the ensemble temperature covariate. Thus, the skewness might need to be included using an appropriate response distribution without increasing the model complexity with additional covariates. Such covariates would also require variable selection techniques to avoid overfitting.

In this study, the skewed logistic distribution was used and compared to the (symmetric) logistic and Gaussian distributions for probabilistic post-processing of the 2 m air temperature at 27 sites in central Europe for stations in three different environments: Alpine, foreland close to the Alps, and sites located in plain regions. The skewed logistic distribution allows one to directly handle possible skewness in the data, if needed.

The two logistic distributions perform better for 1 d up to 4 d ahead forecasts for the majority of the stations and lead times – in particular regarding sharpness and logarithmic score (LS) – without decreasing calibration, which is analyzed by the reliability index (RI) and probability integral transform (PIT) histograms. The amount of improvement decreases with the decreasing complexity of the topography.

When PIT histograms are used to check for calibration, they have to be checked for different seasons, lead times, and hours of day. Averaging over the whole year or multiple times of the day may mask shortcomings especially in complex terrain, and the distinct patterns as shown in the results might easily be overlooked.

A comparison to sliding window models, where a fixed number of previous days is used for training, highlights that the sliding window approach obtains sharp forecasts, but results in uncalibrated forecasts regarding PIT histograms. A longer sliding window of 60 d compared with 30 d decreases the sharpness of the probabilistic forecasts; however, it is still not calibrated and indicates that skewness occurs in the residuals. Consequently, longer training windows would

have even larger issues with residual skewness. To overcome this, the current study uses a long-term training approach of 3 years and accounts for seasonality. This additional seasonality reduces most parts of the skewness, but still improves the sharpness without decreasing calibration.

The sliding training approach has the advantage of being able to react to and account for changes in the ensemble model quickly if two statistically different time periods exist. The long-term approach would need a refitting of the regressions coefficients for the new period after a change occurred, or the change would have to be treated in the statistical models if two periods are mixed during training.

In conclusion, the Gaussian assumption for probabilistic temperature post-processing may be appropriate for regions where the ensemble provides sufficient information regarding the marginal distribution of the response. However, if the covariates used in the regression model miss some features, residual skewness becomes challenging. An alternative response distribution, such as the proposed skewed logistic distribution, allows one to directly address unresolved skewness and increases the predictive performance of the probabilistic forecasts.

Code availability. The results of the models including smooth splines have been achieved using the R package "bamlss" (Umlauf et al., 2018), where a new family for the generalized logistic type I distribution has been implemented and is now available on R-Forge using the distributional properties from the R package "glogis" (Zeileis and Windberger, 2014). The estimation of these models is performed using a gradient boosting approach with a 10-fold cross-validation to find the optimal stopping iteration for the boosting based on the RMSE in order to achieve regularized regression parameters. All models using a sliding window approach are based on the R package "crch" (Messner et al., 2016) employing frequentist maximum likelihood and CRPS optimization.

Appendix A: Skewness of the skewed logistic distribution

The third moment (skewness, ν) is a function of the shape parameter ζ:

$$\nu(\zeta) = \frac{\Psi''(\zeta) - \Psi''(1)}{(\Psi'(\zeta) + \Psi'(1))^{\frac{3}{2}}}, \tag{A1}$$

where Ψ' and Ψ'' denote the first and second derivative of the polygamma function $\Psi(x)$ (Abramowitz and Stegun, 1965, Sect. 6.4.1, p. 260) defined as

$$\Psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}. \tag{A2}$$

Here, $\Gamma(x)$ denotes the Gamma function (Abramowitz and Stegun, 1965, Sect. 6.1.1, p. 255) and $\Gamma'(x)$ is its first derivative. The Gamma function is defined as

$$\Gamma(x) = \int_0^\infty t^{x-1} \exp(-t) \mathrm{d}t. \tag{A3}$$

Author contributions. This study is based on the PhD work of MG under supervision of GJM and AZ. Simulations were performed by MG and RS; this involved a strong effort from RS, who adjusted the BAMLSS framework. Verification and visualization was performed by MG, who also prepared the paper and the initial concept. All authors worked strongly together discussing the results and commented on the paper.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. Results were partly achieved utilizing the high-performance computing infrastructure at the University of Innsbruck using the supercomputer LEO.

Financial support. This project was partially funded by doctoral funding from the University of Innsbruck, Vizerektorat für Forschung, and the Austrian Research Promotion Agency (FFG), project "Prof-Cast" (grant no. 858537).

Review statement. This paper was edited by Dan Cooley and reviewed by Gregory Herman and two anonymous referees.

References

Abramowitz, M. and Stegun, I. A.: Handbook of mathematical functions with formulas, graphs and mathematical tables, National Bureau of Standards Applied Mathematics Series No. 55, J. Appl. Mech., 32, available at: http://people.math.sfu.ca/~cbm/aands/abramowitz_and_stegun.pdf (last access: 13 June 2019), 1965.

Aldrich, J.: R. A. Fisher and the making of maximum likelihood 1912–1922, Stat. Sci., 12, 162–176, https://doi.org/10.1214/ss/1030037906, 1997.

Anderson, J. L.: A method for producing and evaluating probabilistic forecast from ensemble model integration, J. Climate, 9, 1518–1530, https://doi.org/10.1175/1520-0442(1996)009<1518:AMFPAE>2.0.CO;2, 1996.

Bauer, P., Thorpe, A., and Brunet, G.: The quiet revolution of numerical weather prediction, Nature, 525, 47–55, https://doi.org/10.1038/nature14956, 2015.

Dabernig, M., Mayr, G. J., Messner, J. W., and Zeileis, A.: Spatial ensemble post-processing with standardized anomalies, Q. J. Roy. Meteor. Soc., 143, 909–916, https://doi.org/10.1002/qj.2975, 2017.

Dawid, A.: Present position and potential developments: Some personal views: Statistical theory: The prequential approach, J. R. Stat. Soc. Ser. A-G., 147, 278–292, https://doi.org/10.2307/2981683, 1984.

Delle Monache, L., Hacker, J. P., Zhou, Y., Deng, X., and Stull, R. B.: Probabilistic aspects of meteorological and ozone regional ensemble forecasts, J. Geophys. Res., 111, 1–15, https://doi.org/10.1029/2005JD006917, 2006.

Feldmann, K., Scheuerer, M., and Thorarinsdottir, T. L.: Spatial postprocessing of ensemble forecasts for temperature using nonhomogeneous Gaussian regression, Mon. Weather Rev., 143, 955–971, https://doi.org/10.1175/MWR-D-14-00210.1, 2015.

Gebetsberger, M., Messner, J. W., Mayr, G. J., and Zeileis, A.: Fine-tuning non-homogeneous regression for probabilistic precipitation forecasts: Unanimous predictions, heavy tails, and link functions, Mon. Weather Rev., 145, 4693–4708, https://doi.org/10.1175/MWR-D-16-0388.1, 2017.

Gebetsberger, M., Messner, J. W., Mayr, G. J., and Zeileis, A.: Estimation methods for nonhomogeneous regression models: minimum continuous ranked probability score versus maximum likelihood, Mon. Weather Rev., 146, 4323–4338, https://doi.org/10.1175/MWR-D-17-0364.1, 2018.

Gneiting, T. and Katzfuss, M.: Probabilistic forecasting, Annu. Rev. Stat. Appl., 1, 125–151, https://doi.org/10.1146/annurev-statistics-062713-085831, 2014.

Gneiting, T., Raftery, A. E., Westveld, A. H., and Goldman, T.: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation, Mon. Weather Rev., 133, 1098–1118, https://doi.org/10.1175/MWR2904.1, 2005.

Gneiting, T., Balabdaoui, F., and Raftery, A. E.: Probabilistic forecasts, calibration and sharpness, J. R. Stat. Soc. B, 69, 243–268, https://doi.org/10.1111/j.1467-9868.2007.00587.x, 2007.

Hagedorn, R., Hamill, T., and Whitaker, J.: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: two-meter temperatures, Mon. Weather Rev., 136, 2608–2619, https://doi.org/10.1175/2007MWR2410.1, 2008.

Hamill, T. M. and Colucci, S. J.: Evaluation of Eta RSM ensemble probabilistic precipitation forecasts, Mon. Weather Rev., 126, 711–724, https://doi.org/10.1175/1520-0493(1998)126<0711:EOEREP>2.0.CO;2, 1998.

Harmel, R. D., Richardson, C. W., Hanson, C. L., and Johnson, G. L.: Evaluating the adequacy of simulating maximum and minimum daily air temperature with the normal distribution, J. Appl. Meteorol., 41, 744–753, https://doi.org/10.1175/1520-0450(2002)041<0744:Etaosm>2.0.Co;2, 2002.

Hersbach, H.: Decomposition of the continuous ranked probability score for ensemble prediction systems, Weather Forecast., 15, 559–570, https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2, 2000.

Klein, N., Kneib, T., Lang, S., and Sohn, A.: Bayesian structured additive distributional regression with an application to regional income inequality in Germany, Ann. Appl. Stat., 9, 1024–1052, https://doi.org/10.1214/15-AOAS823, 2015.

Leith, C.: Theoretical skill of Monte Carlo forecasts, Mon. Weather Rev., 102, 409–418, https://doi.org/10.1175/1520-0493(1974)102<0409:TSOMCF>2.0.CO;2, 1974.

Lorenz, E. N.: Deterministic nonperiodic flow, J. Atmos. Sci., 20, 130–141, https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2, 1963.

Messner, J. W., Mayr, G. J., Wilks, D. S., and Zeileis, A.: Extending extended logistic regression: Extended versus separate versus ordered versus censored, Mon. Weather Rev., 142, 3003–3014, https://doi.org/10.1175/MWR-D-13-00355.1, 2014.

Messner, J. W., Mayr, G. J., and Zeileis, A.: Heteroscedastic Censored and Truncated Regression with crch, R J., 8, 173–181, 2016.

Messner, J. W., Mayr, G. J., and Zeileis, A.: Nonhomogeneous boosting for predictor selection in ensemble postprocessing, Mon. Weather Rev., 145, 137–147, https://doi.org/10.1175/MWR-D-16-0088.1, 2017.

Möller, A. and Groß, J.: Probabilistic temperature forecasting based on an ensemble autoregressive modification, Q. J. Roy. Meteor. Soc., 142, 1385–1394, https://doi.org/10.1002/qj.2741, 2016.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M.: Using Bayesian model averaging to calibrate forecast ensembles, Mon. Weather Rev., 133, 1155–1174, https://doi.org/10.1175/MWR2906.1, 2005.

Roulston, M. S. and Smith, L. A.: Combining dynamical and statistical ensembles, Tellus, 55A, 16–30, https://doi.org/10.1034/j.1600-0870.2003.201378.x, 2003.

Schefzik, R., Thorarinsdottir, T. L., and Gneiting, T.: Uncertainty quantification in complex simulation models using ensemble copula coupling, Stat. Sci., 28, 616–640, 2013.

Scheuerer, M. and Büermann, L.: Spatially adaptive post-processing of ensemble forecasts for temperature, J. R. Stat. Soc. C-Appl., 63, 405–422, https://doi.org/10.1111/rssc.12040, 2014.

Stauffer, R., Umlauf, N., Messner, J. W., Mayr, G. J., and Zeileis, A.: Ensemble postprocessing of daily precipitation sums over complex terrain using censored high-resolution standardized anomalies, Mon. Weather Rev., 145, 955–969, https://doi.org/10.1175/MWR-D-16-0260.1, 2017.

Stauffer, R., Mayr, G. J., Messner, J. W., and Zeileis, A.: Hourly probabilistic snow forecasts over complex terrain: a hybrid ensemble postprocessing approach, Adv. Stat. Clim. Meteorol. Oceanogr., 4, 65–86, https://doi.org/10.5194/ascmo-4-65-2018, 2018.

Steinacker, R.: Area height distribution of a valley and its relation to the valley wind, Beitr. Phys. Atmos., 57, 64–71, 1984.

Talagrand, O., Vautard, R., and Strauss, B.: Evaluation of probabilistic prediction systems, in: Proceeding of workshop on predictability, European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading, Berkshire RG2 9AX, UK, 1–25, 1997.

Toth, Z. and Szentimrey, T.: The binormal distribution: A distribution for representing asymmetrical but normal-like weather elements, J. Climate, 3, 128–136, https://doi.org/10.1175/1520-0442(1990)003<0128:TBDADF>2.0.CO;2, 1990.

Umlauf, N., Klein, N., and Zeileis, A.: BAMLSS: Bayesian additive models for location, scale and shape (and beyond), J. Comput. Graph. Stat., 27, 612–627, https://doi.org/10.1080/10618600.2017.1407325, 2018.

Verkade, J. S., Brown, J. D., Reggiani, P., and Weerts, H.: Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales, J. Hydrol., 501, 73–91, https://doi.org/10.1016/j.jhydrol.2013.07.039, 2013.

Warwick, G. and Curran, E.: A binormal model of frequency distributions of daily maximum temperature, Aust. Meteorol. Mag., 42, 151–161, 1993.

Whiteman, C.: Observations of thermally developed wind systems in mountainous terrain, Atmospheric processes over complex terrain, Meteor. Mon., 23, 5–42, https://doi.org/10.1007/978-1-935704-25-6_2, 1990.

Wilks, D.: Extending logistic regression to provide full-probability-distribution MOS forecasts, Meteorol. Appl., 368, 361–368, https://doi.org/10.1002/met.134, 2009.

Wilks, D. S.: Statistical methods in the atmospheric sciences, 3 edn., Elsevier Academic Press, Oxford, UK, Amsterdam, the Netherlands, San Diego, Ca, USA, 704 pp., 2011.

Wilks, D. S.: On assessing calibration of multivariate ensemble forecasts, Q. J. Roy. Meteor. Soc., 143, 164–172, https://doi.org/10.1002/qj.2906, 2017.

Zängl, G.: A reexamination of the valley wind system in the Alpine Inn Valley with numerical simulations, Meteorol. Atmos. Phys., 87, 241–256, https://doi.org/10.1007/s00703-003-0056-5, 2004.

Zeileis, A. and Windberger, T.: glogis: Fitting and testing generalized logistic distributions, https://CRAN.R-project.org/package=glogis (last access: 13 June 2019), 2014.

Article VII

Lang M.N., Mayr G.J., Stauffer R., and Zeileis A. (2019). *Bivariate Gaussian Models for Wind Vectors in a Distributional Regression Framework*. Advances in Statistical Climatoloy, Meteorology and Oceanography, 5, 115–132, doi:10.5194/ASCMO-5-115-2019.

Recent peer-reviewed journal on the intersection of atomspheric science and statistics (published by Copernicus), not yet listed in JCR.

Contribution (CRT): *Conceptualization / data curation / formal analysis / validation / supervision (informal role) / writing, original draft*.

Adv. Stat. Clim. Meteorol. Oceanogr., 5, 115–132, 2019
https://doi.org/10.5194/ascmo-5-115-2019

ASCMO
Open Access

Bivariate Gaussian models for wind vectors in a distributional regression framework

Moritz N. Lang[1,2], **Georg J. Mayr**[2], **Reto Stauffer**[1], **and Achim Zeileis**[1]

[1]Department of Statistics, University of Innsbruck, Innsbruck, Austria
[2]Department of Atmospheric and Cryospheric Sciences, University of Innsbruck, Innsbruck, Austria

Correspondence: Moritz N. Lang (moritz.lang@uibk.ac.at)

Received: 2 April 2019 – Revised: 4 June 2019 – Accepted: 11 June 2019 – Published: 18 July 2019

Abstract. A new probabilistic post-processing method for wind vectors is presented in a distributional regression framework employing the bivariate Gaussian distribution. In contrast to previous studies, all parameters of the distribution are simultaneously modeled, namely the location and scale parameters for both wind components and also the correlation coefficient between them employing flexible regression splines. To capture a possible mismatch between the predicted and observed wind direction, ensemble forecasts of both wind components are included using flexible two-dimensional smooth functions. This encompasses a smooth rotation of the wind direction conditional on the season and the forecasted ensemble wind direction.

The performance of the new method is tested for stations located in plains, in mountain foreland, and within an alpine valley, employing ECMWF ensemble forecasts as explanatory variables for all distribution parameters. The rotation-allowing model shows distinct improvements in terms of predictive skill for all sites compared to a baseline model that post-processes each wind component separately. Moreover, different correlation specifications are tested, and small improvements compared to the model setup with no estimated correlation could be found for stations located in alpine valleys.

1 Introduction

Accurate forecasts of wind speed and direction are of great importance for decision-making processes and risk management in today's society and will likely become more important in the future. This is not only because of the rapid change in climate and the resulting increase in severe storms (e.g., Kunkel et al., 2012; Vose et al., 2013), but is also due to the change in the society itself and its technical revolution. As an example, the European Union is aiming to increase the amount of wind energy by 2030 to 35 %, which would be more than double the capacity installed at the end of 2016 (WindEurope, 2017). In the field of aviation and air traffic control for instance, more flexible landing procedures with a so-called time-based separation are currently being tested at Heathrow Airport and are planned to go operational in the near future (EuropeanCommission, 2018). In both cases, wind (power) forecasts are of fundamental importance; probabilistic wind forecasts are in particular advisable as they permit optimal risk assessment and decision making (Gneiting, 2008).

Probabilistic weather forecasts are usually issued in the form of ensemble predictions. To account for the underlying uncertainty in the atmosphere, numerical ensemble prediction systems (EPSs) provide a set of weather forecasts using slightly perturbed initial conditions and different model parameterizations (Palmer, 2002). Despite recent advances in the development of EPSs, the resulting forecasts still often show displacement errors and usually capture only part of the forecast uncertainty, especially when comparing EPS forecasts and point measurements (Buizza et al., 2005; Gneiting and Katzfuss, 2014). This often results from structural model deficiencies and insufficient resolution or unresolved topographical features. To remove systematic biases and to provide corrected variance information, statistical post-processing methods are often employed. For wind, various ensemble post-processing methods have been proposed over the last decade, mainly focusing on

wind speed. For a single location, parametric examples are non-homogeneous regression (Thorarinsdottir and Gneiting, 2010; Lerch and Thorarinsdottir, 2013; Baran and Lerch, 2015, 2016), kernel dressing methods with similarities to Bayesian model averaging (Sloughter et al., 2010; Courtney et al., 2013; Baran, 2014), and extended logistic regression (Messner et al., 2014a, b). A non-parametric approach based on quantile regression forests was applied by Taillardat et al. (2016). On a regular grid, ensemble post-processing based on non-homogeneous regression was performed by Scheuerer and Möller (2015).

To account for the circular characteristics of wind or utilizing information of wind speed and direction, an intuitive post-processing approach is to model a bivariate process for the zonal and meridional wind components. Gneiting et al. (2008) suggested using a bivariate Gaussian response distribution for the wind components, an idea that was implemented by Pinson (2012). He estimates a dilation and translation factor for the individual ensemble members utilizing the empirical correlation structure of the EPS. This procedure can be seen as a variant of the ensemble copula coupling (ECC) method introduced by Schefzik et al. (2013). With the ECC, both wind components are calibrated with univariate approaches and a discrete sample drawn from each univariate predictive distribution is rearranged in the rank order structure of the raw ensemble. The method introduced by Schuhen et al. (2012) also fits a bivariate Gaussian distribution for the wind components; however, in their approach the post-processed probabilistic forecast consists of a fully specified predictive distribution instead of a discrete ensemble. As their analyses show that the observed correlation between the two wind components mainly depends on wind direction, they model the correlation parameter in the bivariate Gaussian distribution as a trigonometric function of the ensemble mean wind direction. An extra group is formed for cases with low wind speeds unconditionally on their wind direction. The estimation of the correlation parameter is done offline in a pre-processing step for a separate year, either for all stations combined or for each station individually. However, according to Schuhen et al. (2012), the fitting can be critical for individual stations since wind sectors may only contain a few data points.

For stations in complex terrain, a possible drawback of the bivariate post-processing approach of Schuhen et al. (2012) is that the model cannot correct for a systematic distortion in the wind directions due to discrepancies between the model and real topography. Especially when the respective valley orientations differ, a meridional wind component might be partially rotated into a zonal wind component and vice versa. Pinson (2012) employs both forecasted wind components for the calibration of the zonal or meridional wind component, which is partly able to correct for systematic distortions in wind directions in a linear manner. In the field of post-processing deterministic weather forecasts, this approach was already suggested by Glahn and Lowry (1972).

Alternatively to bivariate calibration methods, wind direction can also be employed in univariate settings. In a post-processing approach for wind speed, Eide et al. (2017) suggest utilizing the potentially nonlinear information of the wind direction by a generalized additive model (GAM; Hastie and Tibshirani, 1986). GAMs were first applied in the meteorological context by Vislocky and Fritsch (1995) and provide a powerful statistical model framework which can capture potential nonlinear relationships between the covariates and the response by smooth functions or splines. Eide et al. (2017) employ wind direction as an additional covariate for the estimation of wind speed, by accounting for its cyclic and potential nonlinear characteristics utilizing thin-plate regression splines.

In this study, we directly model the zonal and meridional wind components, employing the bivariate Gaussian distribution as suggested by Gneiting et al. (2008) and performed by Pinson (2012) and Schuhen et al. (2012). However, we capture all distribution parameters, namely the location and scale parameters for both wind components, and also the correlation coefficient between them in a single flexible model. In the estimation of the two-dimensional location and scale parameters the information value of both ensemble wind components is utilized to allow for a smooth rotation of the forecasted wind direction accounting for unresolved topographical features. To consider the correlation characteristics detected in Schuhen et al. (2012) and to allow for possible nonlinear effects, such as, e.g., suggested by Eide et al. (2017), we model the correlation as a function of wind speed and direction utilizing cyclic regression splines. To account for potential time-varying effects, all linear predictors use a time-adaptive intercept and time-adaptive slope coefficients based on cyclic smooth splines.

The paper is structured as follows: Sect. 2 introduces the employed statistical models. The underlying data of this study are shortly described in Sect. 3. Within Sect. 4, first a model comparison and validation are presented for two weather stations with different site characteristics, followed by aggregated scores for station sites located in plains, in mountain foreland, and within an alpine valley. The article ends with a brief discussion and a conclusion given in Sect. 5.

2 Methods

In Sect. 2.1, the bivariate Gaussian distribution is reviewed and briefly presented in a distributional regression framework. Subsequently, three broad model classes are introduced, all of which are based on a time-adaptive training scheme but employ different specifications for the location, scale, and correlation parameters of the bivariate distribution. First, the baseline model is presented in Sect. 2.2 that serves as a benchmark and simply combines two univariate heteroscedastic regression models that post-process each wind component separately. Second, the baseline model is

extended in Sect. 2.3 by always adding both EPS wind components as regressors using smooth splines and thus allowing for potential misspecifications in the EPS wind direction. Finally, Sect. 2.4 also considers models with estimated correlation coefficients based on various regression specifications. Table 1 provides a synoptic summary of all bivariate Gaussian model specifications tested within this study.

2.1 Distributional regression for a bivariate Gaussian response

The zonal and meridional components of the horizontal wind vector are represented by a bivariate Gaussian distribution. Its likelihood function L is given by

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \boldsymbol{y}) = \frac{1}{\sqrt{(2\pi)^2 |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right), \quad (1)$$

where $\boldsymbol{y} = (y_1, y_2)^\top$ are bivariate observations and $\boldsymbol{\mu} = (\mu_1, \mu_2)^\top$ are the distributional location parameters with $\mu_\star \in \mathbb{R}$; the subscript asterisk acts as a placeholder for the zonal and meridional wind components from here on. The covariance matrix is defined as

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}, \quad (2)$$

with correlation parameter $\rho \in [-1, 1]$ and scale parameters $\sigma_\star > 0$. In the framework of distributional regression, link functions provide the relationship between unrestricted linear predictors and the respective distribution parameters by ensuring their appropriate co-domain. For the bivariate Gaussian distribution, the location parameters $\mu_\star$, the scale parameters $\sigma_\star$, and the correlation parameter ρ are linked to their additive predictors by an identity, logarithm and rhogit link, respectively (Klein et al., 2014).

To be able to utilize the information of cyclic covariates, such as, e.g., wind direction in addition to linear covariates, we follow Eide et al. (2017) and fit a GAM but utilize cubic smooth functions with cyclic constraints for all cyclic covariates (Wood, 2017; see Appendix A1). In the context of distributional regression, additive models with smooth effects are typically referred to as "generalized additive models for location, scale and shape" (GAMLSS, Rigby and Stasinopoulos, 2005). In this study, we utilize GAMLSS in a Bayesian framework, which allows us to examine the estimated effects based on Markov chain Monte Carlo (MCMC) simulations and ensures stable estimation of the regression coefficients. The key steps involved in the estimation can be summarized as follows. The parameters of the bivariate Gaussian distribution are linked to the set of additive predictors containing potentially nonlinear transformations of the covariates according to the model specifications described later on in this section. The model fitting is performed by a derivative-based MCMC sampling using iteratively weighted least squares (IWLS, Gamerman, 1997) proposals. Hence, the estimated effects are based on 1000 independent realizations of the regression coefficients drawn from the Markov chains and for subsequent predictions the means of these samples are used as point estimators for the regression coefficients. A comprehensive summary of the method is given in Umlauf et al. (2018).

2.2 Baseline model (BLM-0)

The baseline model (BLM-0) combines two univariate heteroscedastic regression models that post-process each wind component separately with correlation fixed at zero. Hence, for the location and scale part, it uses its direct counterparts of the EPS as covariates, namely EPS-forecasted zonal wind information (vec_1) to model the zonal component of the bivariate response and EPS-forecasted meridional wind information (vec_2) to model the meridional component:

$$\begin{aligned} \mu_\star &= \underbrace{\alpha_{\star 0} + f_{\star 0}(\text{doy})}_{\text{intercept}} + \underbrace{(\alpha_{\star 1} + f_{\star 1}(\text{doy}))}_{\text{slope coefficient}} \cdot \text{vec}_{\star,\text{mean}}, \\ \log(\sigma_\star) &= \underbrace{\beta_{\star 0} + g_{\star 0}(\text{doy})}_{\text{intercept}} + \underbrace{(\beta_{\star 1} + g_{\star 1}(\text{doy}))}_{\text{slope coefficient}} \cdot \text{vec}_{\star,\text{log.sd}}, \end{aligned} \quad (3)$$

where $\alpha_\bullet$ and $\beta_\bullet$ are regression coefficients, and $f_\bullet(\text{doy})$ and $g_\bullet(\text{doy})$ employ cyclic regression splines conditional on the day of the year (doy). The subscripts mean and log.sd refer to the mean and log standard deviation of the ensemble wind components, respectively. We follow Gebetsberger et al. (2017) and use the logarithm transformation for the standard deviation of the ensemble members to ensure positivity, which is preferable for the estimation process.

Equation (3) specifies a time-adaptive training scheme (with further details in Appendix A2), where the linear predictors consist of a global intercept and slope coefficient plus a seasonally varying deviation. Thus, the intercept and slope coefficients can smoothly evolve over the year in case the bias or the covariate's skill varies seasonally. If there is no seasonal variation, the nonlinear effects become zero and Eq. (3) simplifies to a regression model with a constant intercept and slope coefficient ($\mu_\star = \alpha_{\star 0} + \alpha_{\star 1} \cdot \text{vec}_{\star,\text{mean}}$; $\log(\sigma_\star) = \beta_{\star 0} + \beta_{\star 1} \cdot \text{vec}_{\star,\text{log.sd}}$).

2.3 Rotation-allowing model without correlation (RAM-0)

In the second model, labeled the rotation-allowing model (RAM-0), we extend the BLM-0 setup by employing the zonal and meridional wind information of the ensemble for the linear predictors of all location and scale parameters. That means we use the ensemble information of both the zonal and meridional wind components for the two components of the response (cf. Glahn and Lowry, 1972). In case of a perfect EPS the zonal wind predictions are noninformative covariates for the meridional wind component and vice versa. However, if, e.g., the model topography is

Table 1. Overview of bivariate Gaussian model specifications. For the "baseline model" (BLM-0; see Sect. 2.2) and the "rotation-allowing model" (RAM-0; see Sect. 2.3) no correlation is employed, i.e., fixed at zero. For all tested correlation specifications, the RAM-0 setup is employed for the location and scale part (see Sect. 2.4). In all setups for each distribution parameter, a seasonally varying intercept effect is estimated. For BLM-0, a seasonally varying slope coefficient is fitted for the two wind components in the location and scale parts. For the RAM-0 setup, the slope coefficients are additionally dependent on the wind direction. In the RAM-ADV correlation model, the wind speed is modeled conditional on the wind direction. The equal sign expresses "the response is set to" and the tilde signals "the response is modeled by" the term(s) on the right-hand side of the equation. The symbol "– " –" implies that the same configuration as in the line above is employed.

Name	Location part	Scale part	Correlation part
BLM-0	$\mu_\star \sim \text{vec}_{\star,\text{mean}}$	$\sigma_\star \sim \text{vec}_{\star,\text{log.sd}}$	$\rho = 0$
RAM-0	$\mu_\star \sim \text{vec}_{1,\text{mean}}, \text{vec}_{2,\text{mean}}$	$\sigma_\star \sim \text{vec}_{1,\text{log.sd}}, \text{vec}_{2,\text{log.sd}}$	$\rho = 0$
RAM-EMP	– " –	– " –	$\rho = \text{vec}_{\star,\text{corr}}$
RAM-IC	– " –	– " –	$\rho \sim 1$
RAM-DIR	– " –	– " –	$\rho \sim \text{dir}_{\text{mean}}$
RAM-ADV	– " –	– " –	$\rho \sim \text{dir}_{\text{mean}}, \text{spd}_{\text{mean}}$

not sufficiently resolved or in the case of local shadowing effects, both EPS wind components may contain valuable information for the zonal and meridional wind components of the response. Especially in a mountain valley, when the model and real valley orientation differs, both wind components of the ensemble can potentially contain information about both location and scale parameters, respectively. Thus, we propose to employ seasonally varying effects depending on the ensemble wind direction, which allows the model to rotate the forecasted wind direction if necessary. To do so, we obtain a two-dimensional smooth function represented by a tensor product spline with a respective cyclic constraint for the day of the year (doy) and for the mean ensemble wind direction (dir_{mean}):

$$\begin{aligned}
\mu_\star &= \alpha_{\star 0} + f_{\star 0}(\text{doy}) \\
&\quad + \left(\alpha_{\star 1} + f_{\star 1}(\text{doy}) \cdot f_{\star 2}(\text{dir}_{\text{mean}})\right) \cdot \text{vec}_{1,\text{mean}} \\
&\quad + \left(\alpha_{\star 2} + f_{\star 3}(\text{doy}) \cdot f_{\star 4}(\text{dir}_{\text{mean}})\right) \cdot \text{vec}_{2,\text{mean}}, \\
\log(\sigma_\star) &= \beta_{\star 0} + g_{\star 0}(\text{doy}) \\
&\quad + \left(\beta_{\star 1} + g_{\star 1}(\text{doy}) \cdot g_{\star 2}(\text{dir}_{\text{mean}})\right) \cdot \text{vec}_{1,\text{log.sd}} \\
&\quad + \left(\beta_{\star 2} + g_{\star 3}(\text{doy}) \cdot g_{\star 4}(\text{dir}_{\text{mean}})\right) \cdot \text{vec}_{2,\text{log.sd}}, \qquad (4)
\end{aligned}$$

where, as before, $\alpha_\bullet$ and $\beta_\bullet$ are regression coefficients, and $f_\bullet$ and $g_\bullet$ employ cyclic regression splines. From a more physical perspective, the two-dimensional smooth effects rotate the ensemble wind components conditional on the day of the year and the ensemble wind direction.

2.4 Rotation-allowing models with correlation

By explicitly modeling the correlation, we further extend the RAM-0 setup within this section. For the estimation of the correlation structure different model specifications are tested. The most advanced specification, RAM-ADV, assumes that the correlation mainly depends on the mean ensemble wind direction (dir_{mean}) and speed (spd_{mean}) by modeling a linear interaction between these two covariates:

$$\begin{aligned}
\text{rhogit}(\rho) &= \gamma_0 + h_0(\text{doy}) + h_1(\text{dir}_{\text{mean}}) \\
&\quad + (\gamma_1 + h_2(\text{dir}_{\text{mean}})) \cdot \text{spd}_{\text{mean}}, \qquad (5)
\end{aligned}$$

with $\text{rhogit}(\rho) = \rho/\sqrt{(1-\rho^2)}$; γ_0 is the global intercept and $h_0(\text{doy})$ the seasonally varying intercept. The effect $h_1(\text{dir}_{\text{mean}})$ estimates the dependence of the correlation given the wind direction and $(\gamma_1 + h_2(\text{dir}_{\text{mean}})) \cdot \text{spd}_{\text{mean}}$ employs a varying effect of wind speed conditional on the wind direction. The estimation of the underlying correlation structure is in accordance with results of Schuhen et al. (2012), who employ wind direction and an offset of wind speed as informative covariates in the estimation of the correlation parameter.

Other implementations tested for the correlation parameter are an intercept-only model (RAM-IC), a model with a cyclic effect solely depending on wind direction (RAM-DIR), the RAM-0 independent-component model (Sect. 2.3), and a model using the empirical correlation (corr) of the raw ensemble (RAM-EMP). A synoptic table of all models tested in this study is given in Table 1.

3 Data

3.1 Observational data

The validation and comparison of the different model specifications are performed for 15 measurement sites located across Austria, Germany, and Switzerland. The sites are chosen to investigate the influence of different underlying topographies or varying discrepancies between the real and model topography on the post-processing. The stations are divided into three groups representing sites located in plains, mountain foreland, and within an alpine valley. An overview of the stations is given in Fig. 1. The results for stations Hamburg and Innsbruck, which are labeled in Fig. 1, are discussed in more detail in Sect. 4. At all meteorological sites, wind speed and direction measurements are reported for the

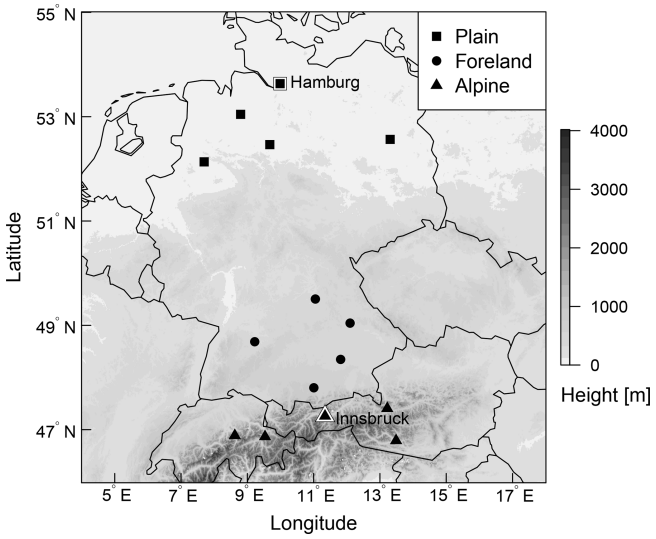

Figure 1. Overview of the study area with selected stations classified as plain, foreland, and alpine station sites. The labeled stations with a white background, Hamburg and Innsbruck, are discussed in detail in Sect. 4. Elevation data are obtained from the SRTM-30 m digital elevation model (NASA JPL, 2013).

10 m height level. The data are 10 min averages for the period 26 January 2010 to 7 March 2016, yielding a total of 2233 d.

3.2 Ensemble prediction system

Covariates are derived from the global 50-member EPS of the European Centre for Medium-Range Weather Forecasts (ECMWF). These EPS forecasts have a horizontal resolution of approximately 30 km (T639) for the time between January 2010 and March 2016 and are bilinearly interpolated to the measurement sites. Covariates employed in this study are the zonal and meridional wind components as well as the derived quantities wind speed and direction valid at 10 m above ground. For all these variables, two statistics are computed over the 50 perturbed ensemble members, namely the mean and the logarithm of the standard deviation (log.sd). Additionally, the Pearson sample correlation coefficient (corr) is computed from the raw ensemble members to capture the correlation between the two wind components. Forecasts are taken from the EPS run initialized at 00:00 UTC for forecast steps ranging from +12 to +72 h ahead at a 12-hourly temporal resolution. Figure 2 shows the empirical wind distributions of the observed and predicted winds for Innsbruck and Hamburg for forecast steps +12 and +24 h corresponding to 12:00 and 00:00 UTC.

4 Results

This section presents the results of the statistical post-processing models. The structure is as follows. First, the estimated effects of the baseline model, BLM-0 (Sect. 4.1), and the rotation-allowing model, RAM-0 (Sect. 4.2), are shown for two stations representative of one alpine valley site and of a measurement site in the plains. For both models a constant correlation of zero is employed, and their predictive performance is discussed in Sect. 4.3. Afterwards, model comparison (Sect. 4.4) and validation (Sect. 4.5) of the different correlation specifications are given for the two representative stations. In Sect. 4.6, the overall performance of the model setups is evaluated for three groups of stations classified as topographically plain, mountain foreland, and alpine valley sites.

The model estimation is performed on data of the first 4 years, leaving an out-of-sample validation data set ranging from 24 February 2014 to 7 March 2016.

4.1 Baseline model (BLM-0)

For BLM-0, the cyclic seasonal effects for stations Hamburg and Innsbruck are shown in Fig. 3 as solid and dashed lines with the respective 95 % credible intervals. The estimated effects are on the scale of the additive predictor, i.e., on the linear scale for the location parameters $\mu_\star$ and on the log scale for the scale parameters $\sigma_\star$. Each of the four distribution parameters is described by a (potentially) seasonally varying effect for the intercept (panels a, c, e and g) and the slope coefficient (panels b, d, f and h) as specified in Eq. (3).

For Hamburg, for both location parameters $\mu_\star$, the intercept effect is almost zero (Fig. 3a, c) and the effect for the slope coefficient is close to one (Fig. 3b, d) with very little seasonal variability. This means apparently no bias correction is necessary and the ensemble mean wind components are mapped almost one-to-one to the location parameters. Similarly, barely any seasonal variation exists for the scale parameters $\sigma_\star$ (Fig. 3e–h); however, here the intercept and slope coefficients actually post-process the EPS variances of the wind components (rather than a one-to-one mapping only), leading to an increase in the scale parameters compared to the under-dispersed ensemble. The 95 % credible intervals indicate a higher uncertainty of the estimated scale parameters compared to the location parameters. In summary, the EPS performance for Hamburg is almost constant over the year, and no time-adaptive training scheme seems to be necessary.

By contrast, for Innsbruck the estimated effects show a distinct annual cycle for the location parameters $\mu_\star$, which indicates a varying information content of the predictor variables $\text{vec}_{\star,\text{mean}}$ and the need for some adaptive training scheme. For the location parameter μ_1, the intercept is rather large during winter (Fig. 3a), while, at the same time, the slope coefficient (Fig. 3b) is close to zero due to an apparently low skill of the EPS. For the location parameter μ_2 (Fig. 3c, d), the higher slope coefficients during spring and autumn suggest a higher information content of the raw EPS in the transitional seasons than for the rest of the year. For the scale parameters (Fig. 3e–h), the estimated effects show high vari-

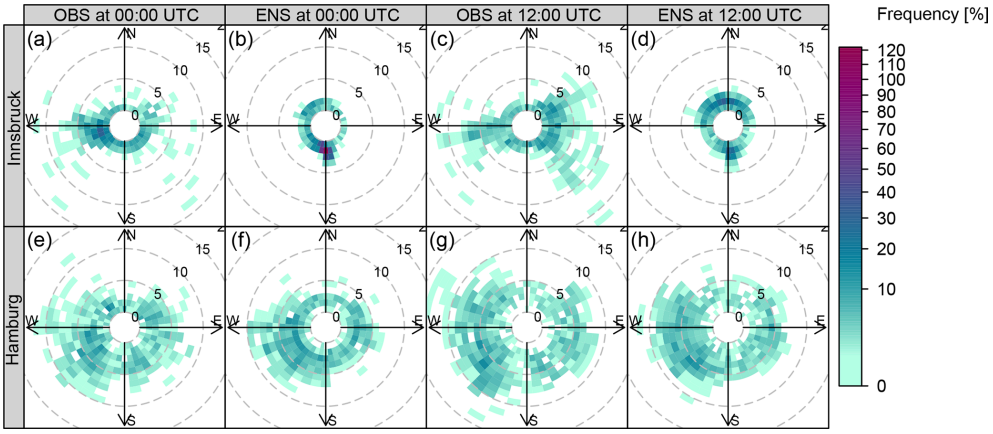

Figure 2. Empirical wind distributions of observations (OBS) and mean ensemble forecasts (ENS) for Innsbruck and Hamburg. The probability of occurrence is color-coded and the wind speed is represented by contour lines ($m\,s^{-1}$). The forecast steps +12 and +24 h, valid at 12:00 and 00:00 UTC, are shown for the validation period from 24 February 2014 to 7 March 2016.

ability; this indicates a seasonally varying skill of the EPS variance information. In summary, for the weather station in Innsbruck, the information content of the ensemble wind predictions seems to be rather low. This is in accordance with the clearly different pattern of observations and EPS forecasts shown in Fig. 2.

4.2 Rotation-allowing model without correlation (RAM-0)

Figure 4 shows the estimated mean effects of the RAM-0 setup in comparison with BLM-0 on the wind direction at Innsbruck and Hamburg for the forecast steps +12 and +24 h valid at 12:00 and 00:00 UTC, respectively. The marginal effects are non-centered and shown for the mean covariates within 10° wind sectors conditional on the day of the year. The BLM-0 model for Innsbruck shows a distinct seasonal dependency of the post-processed wind direction for both times of the day (Fig. 4a, c). During winter at 12:00 and 00:00 UTC, mainly down-valley winds (approximately 280°) are predicted, whereas over the rest of the year, the EPS mainly forecasts up-valley wind directions. This pattern is more pronounced during night (00:00 UTC) and has less variability in summer. In general, the BLM-0 setup seems to mainly capture the climatological mean wind direction; this leads to little variations between the different wind directions issued by the EPS. In contrast, the rotation-allowing RAM-0 setup has the flexibility to post-process the wind directions conditional on the forecasted EPS wind direction, which is apparent for the Innsbruck station at both 12:00 and 00:00 UTC: for 12:00 UTC the seasonal dependency leads to either up-valley or down-valley wind conditional on the ensemble wind direction and on the day of the year (Fig. 4b), whereas at 00:00 UTC almost no seasonal variation exists and the predicted wind direction solely depends on the issued ensemble wind forecasts (Fig. 4d). At Hamburg, a completely different picture can be seen: almost no post-processing conditional on the ensemble wind direction or the day of the year is visible for both time steps and model setups (Fig. 4e–f). In other words, the predicted ensemble wind direction fits the observed wind direction quite well, and only little statistical correction is needed.

4.3 Predictive performance – models without correlation

To investigate the predictive performance of the two competing setups, Fig. 5 shows the discretized logarithmic score based on the bivariate Gaussian distribution (LS; see Appendix B) and the energy score (ES; see Appendix B) for the forecast steps from +12 to +72 h at a 12-hourly temporal resolution. In addition, skill scores are shown with the raw EPS as a reference or comparing the different setups to each other. Both multivariate scores are proper scores (Gneiting and Raftery, 2007) and evaluate the full predictive distribution returned by the statistical models. The scores for the different forecast horizons show an overall better predictive performance at Hamburg than at Innsbruck. For both stations, the forecasts valid at 00:00 UTC have more skill than those for 12:00 UTC, with higher diurnal variations at Innsbruck. In terms of the ES, the improvements of the BLM-0 model over the raw EPS are about 29 % for Innsbruck and 8 % for Hamburg (Fig. 5e, f). In terms of the LS (Fig. 5b, c), the skill scores are higher, with improvements of approximately 87 % and 33 % for Innsbruck and Hamburg, respectively. The predictive performance gain for the more flexible rotation-allowing RAM-0 setup compared to the BLM-0 specification is around 7 % for Innsbruck and 2 % for Hamburg in terms of the ES (Fig. 5e, f). The LS shows slightly less pronounced relative improvements for the more flexible setup (4 % and 1 %; Fig. 5b, c). The distinct improvements in the scores for RAM-0 are as expected for Innsbruck due to a more flexible utilization of the ensemble information. For plain areas like Hamburg, we assume the better performance is based on

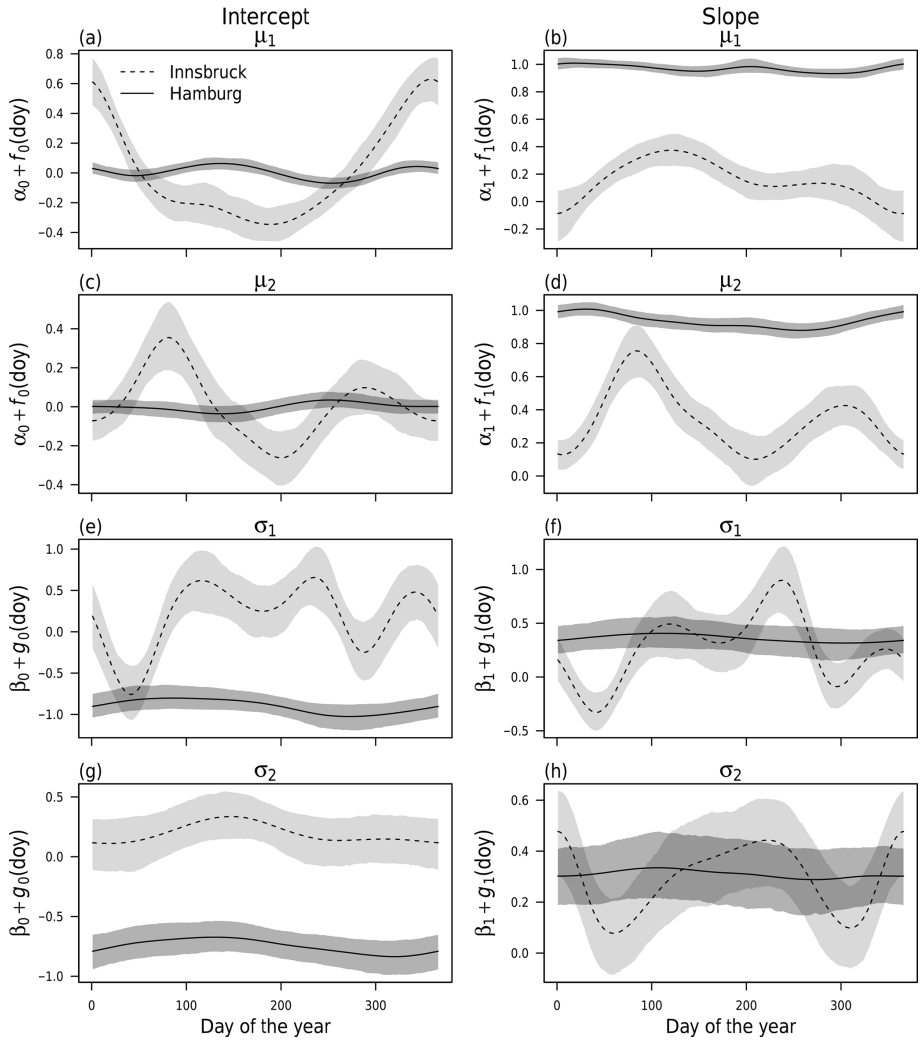

Figure 3. Cyclic seasonal intercept and slope effects according to Eq. (3) employing a constant correlation of zero for weather stations Innsbruck (dashed) and Hamburg (solid) at forecast step +12 h (valid at 12:00 UTC). The effects for the location parameters $\mu_\star$ **(a–d)** and scale parameters $\sigma_\star$ **(e–h)** are shown on the linear and log scales, respectively. The shading represents the 95 % credible intervals based on MCMC sampling.

an enhanced adjustment of the location parameters as both wind components are included in the linear predictors and is not due to the smooth rotation (cf. Fig. 4).

4.4 Rotation-allowing models with correlation

After investigating the two competing location or scale setups, we now focus on an extension of the RAM-0 model by explicitly estimating the underlying correlation structure. Different model specifications for the correlation parameter ρ are tested employing the same linear predictors for $\mu_\star$ and $\sigma_\star$ (see Table 1).

Figure 6 shows correlation parameters predicted by different models for the forecast step +12 h for the full validation period. For comparison, the underlying correlation structure of the raw EPS is also shown. The latter is distributed similarly for Innsbruck and Hamburg and has almost the shape of a Gaussian distribution (Fig. 6a, e). The RAM-IC intercept-only model, with a varying intercept over the year, estimates correlations between −0.27 and 0.48 for Innsbruck (Fig. 6b) and values near zero without clear seasonal variations for Hamburg (Fig. 6f). At both stations, the models with varying effects conditional on the wind direction have similarly distributed correlation parameters with a slightly larger range of predicted values for the RAM-ADV model (Fig. 6d, f) than for the RAM-DIR setup (Fig. 6c, g). The predicted correlation parameters are on average larger for Innsbruck than for Hamburg.

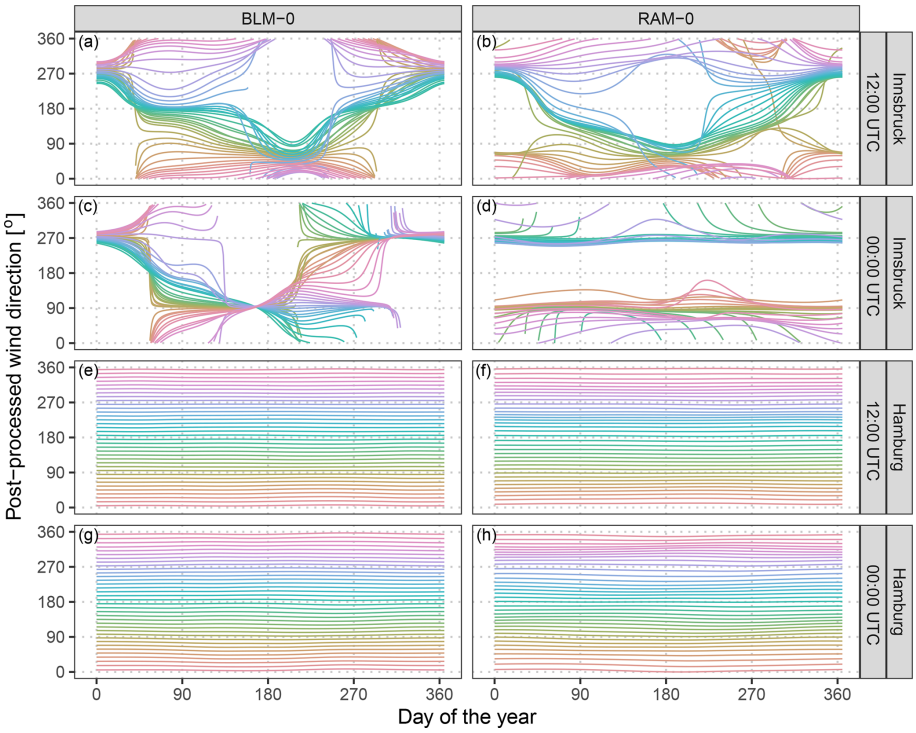

Figure 4. Estimated mean effects for the derived post-processed wind direction at Innsbruck **(a–d)** and Hamburg **(e–h)** for the forecast steps +12 and +24 h (valid at 12:00 and 00:00 UTC). The colored lines show marginal effects for the post-processed wind direction conditional on mean values within 10° wide wind sectors given the training data set. The effects are non-centered and are calculated conditional on the day of the year according to model setups BLM-0 (**a, c, e, g**; Eq. 3) and RAM-0 (**b, d, f, h**; Eq. 4).

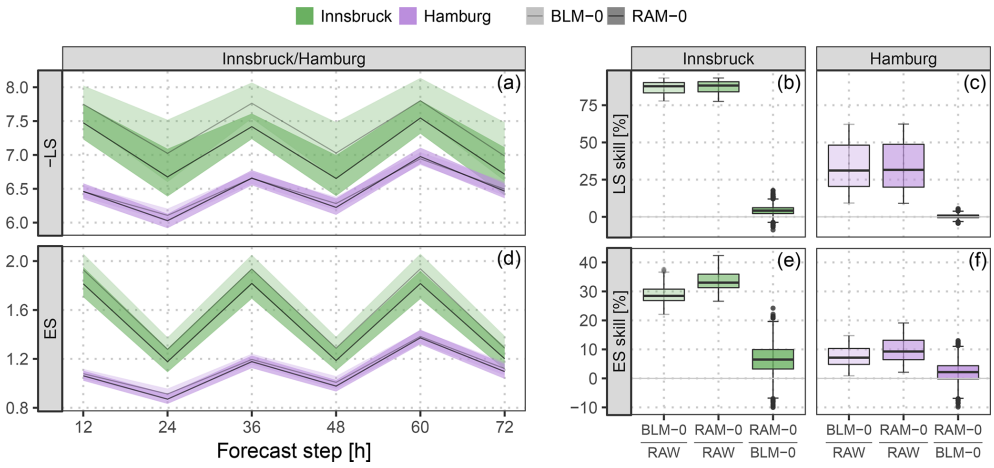

Figure 5. Predictive performance in terms of the logarithmic score (LS) and the energy score (ES) based on the full predictive bivariate distribution for the out-of-sample validation period. The two specifications BLM-0 (Eq. 3) and RAM-0 (Eq. 4) are compared. **(a, d)** Evolution over time for the forecast steps from +12 to +72 h at a 12-hourly temporal resolution. The solid lines represent the mean values per forecast step, the shading the respective 95 % confidence intervals based on boot-strapped mean values. To unify the orientation of both scores, the negative LS is shown (i.e., smaller is better). **(b, c, e, f)** Aggregated skill scores over the forecast steps, comparing the specifications BLM-0 and RAM-0 either against the raw ensemble (RAW) or against each other. Skill scores are in percent; positive values indicate improvements over the reference.

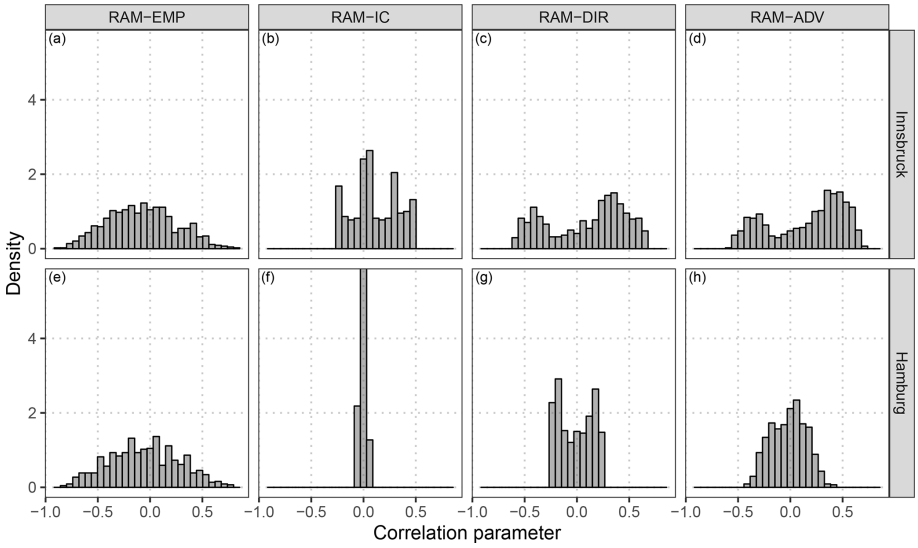

Figure 6. Distribution of the correlation parameters for the underlying dependence structure of the raw ensemble and for the fitted correlation according to the models specified in Table 1. The distributions are shown for Innsbruck **(a–d)** and Hamburg **(e–h)** at the forecast step +12 h for the out-of-sample validation period.

4.5 Predictive performance – models with correlation

Figure 7 shows the verification of bivariate wind speed predictions with an explicitly estimated correlation parameter for Innsbruck and Hamburg; the scores are aggregated over the forecast steps +12 to +72 h at a 12-hourly temporal resolution. As in Fig. 5, the predictive performance is validated in terms of the LS and the ES, based on the full predictive bivariate distributions. However, for comparing different predictive distributions with different correlation structures, the ES' discriminatory ability is limited as it mainly focuses on the location part and hardly discriminates between different correlation structures (Pinson and Tastu, 2013). In Fig. 7, skill scores are shown for the different correlation models with the RAM-0 post-processed model as a reference. The RAM-EMP model, employing the empirical correlation of the raw EPS, performs slightly worse than the reference model for both stations and both scores. This indicates that the raw dependence structure of the EPS has rather low skill. However, for all other models which explicitly model the correlation, only little additional improvement in terms of the LS and the ES is found. At Innsbruck, the RAM-IC intercept-only model performs best in terms of the ES (Fig. 7c). Regarding the LS, minor benefits are present for the most flexible model setup, RAM-ADV (Fig. 7a). For Hamburg, a similar picture is depicted in terms of the LS (Fig. 7b). For the ES (Fig. 7d), the RAM-IC model performs slightly worse than the reference model, and the RAM-ADV model setup performs best.

To validate the calibration of the post-processed predictions, multivariate rank histograms (Gneiting, 2008) are exemplarily shown for the model with no correlation RAM-0 and for the model with the most flexible regression splines in comparison to the raw EPS (Fig. 8). Although the lat-

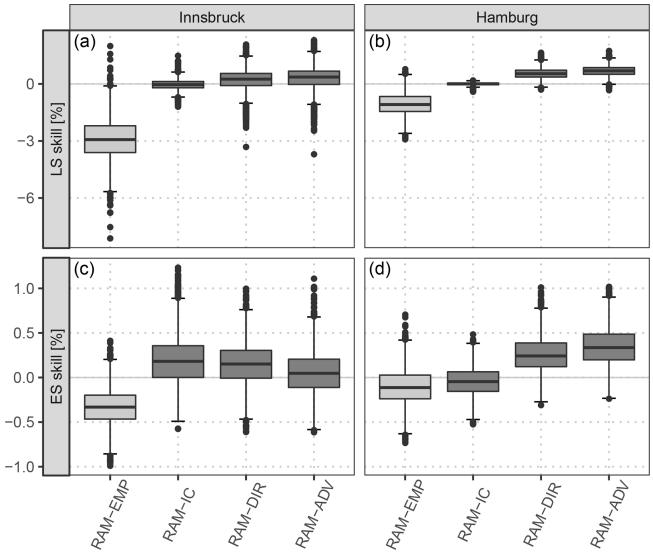

Figure 7. Skill scores aggregated over all forecast steps from +12 to +72 h at a 12-hourly temporal resolution based on the full predictive bivariate distribution for the out-of-sample validation period for Innsbruck **(a, c)** and Hamburg **(b, d)**. Each box-whisker contains boot-strapped mean values per forecast step. The scores are shown for the different correlation models specified in Table 1, with the univariate post-processed model assuming a constant correlation of zero (RAM-0) as a reference. The lighter gray color for the RAM-EMP model indicates that it uses the correlation structure of the raw ensemble without further correlation. Skill scores are in percent; positive values indicate improvements over the reference.

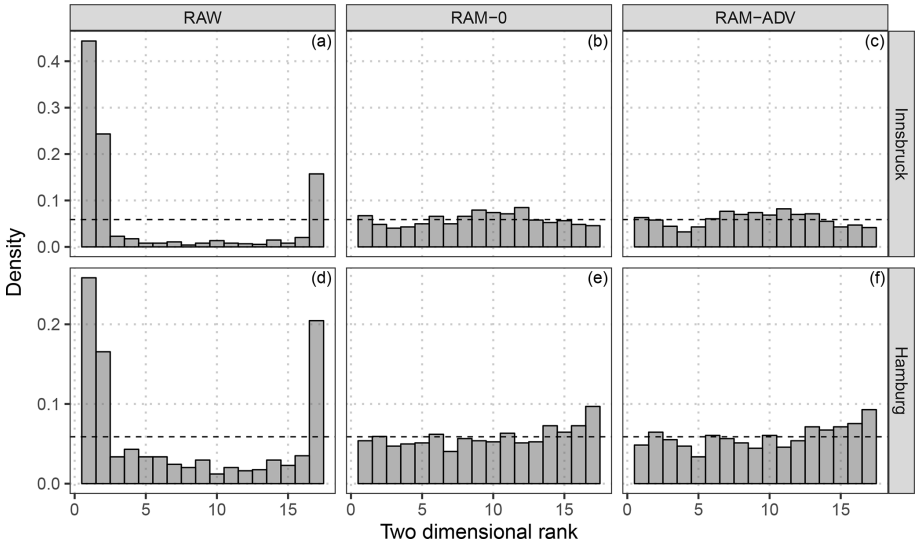

Figure 8. Multivariate rank histograms for raw and post-processed ensemble forecasts according to the correlation model setups RAM-0 and RAM-ADV. The results are shown for Innsbruck **(a–c)** and Hamburg **(d–f)** at the forecast step +12 h for the out-of-sample validation period. For purposes of presentation, three ranks of the raw EPS are combined in a single bar. To stabilize the randomness of ties in the calculation of the multivariate ranks, the median of 20 independent repetitions is plotted.

ter is valid for the grid cell rather than for a single location, both model setups tested are clearly better calibrated than the highly under-dispersive raw ensemble. However, for Innsbruck the multivariate rank histograms of the post-processed forecasts are slightly over-dispersive (Fig. 8b, c) and for Hamburg slightly negatively skewed (Fig. 8e, f). The RAM-ADV flexible model setup shows no significant difference compared to the model with assumed zero correlation (RAM-0).

4.6 Evaluation for all sites

After the previous model comparison at two weather stations, Fig. 9 shows aggregated skill scores for groups of the respective five stations classified as topographically plain, mountain foreland, and alpine valley sites (see Fig. 1). For the location or scale models, two comparisons are shown: the BLM-0 model is compared to the raw EPS as a reference (Fig. 9a, d), and the more flexible rotation-allowing setup, RAM-0, is compared to BLM-0 (Fig. 9b, e). For the correlation specification, the most flexible model, RAM-ADV, is compared to the RAM-0 correlation model employing a constant correlation of zero (Fig. 9c, f).

The post-processing employed by the simplest model, BLM-0, already shows a distinct improvement over the raw EPS with the largest values for alpine valley sites. In terms of the ES, the skill scores range between mean values of 10 % for the plain sites and 45 % for the alpine valley sites (Fig. 9d). A similar picture with an overall larger magnitude is shown for LS (Fig. 9a). In the comparison of the two different setups for the location or scale part (Fig. 9b, e), the more flexible setup is better regarding both scores for all station types; the largest improvements are found for stations located in the foreland, followed by stations within alpine valleys. The validation of the correlation models (Fig. 9c, f) shows that the flexible estimation of the correlation dependence structure is clearly superior only for station sites within an alpine valley.

5 Discussion and conclusion

In this study, we model the zonal and meridional wind components employing the bivariate Gaussian distribution in a distributional regression framework. In contrast to previous studies all distribution parameters, namely the location and scale parameters for both wind components but also the correlation coefficient between them, are estimated simultaneously. The overall performance of the models is evaluated for three groups of station types classified as topographically plain, mountain foreland, and alpine valley sites.

Section 5.1 discusses the benefits of the rotation-allowing model setup, RAM-0, over the baseline model, BLM-0. In Sect. 5.2, the different correlation models are discussed regarding the potential reason why the improvement of the predictive performance obtained with the more flexible correlation model is relatively small. At the end, in Sect. 5.3, a recommendation is given for which statistical model should be used in matters of simplicity and performance.

5.1 Rotation-allowing model setup

The rotation-allowing model (RAM-0) utilizes the zonal and meridional ensemble wind forecasts for both components of

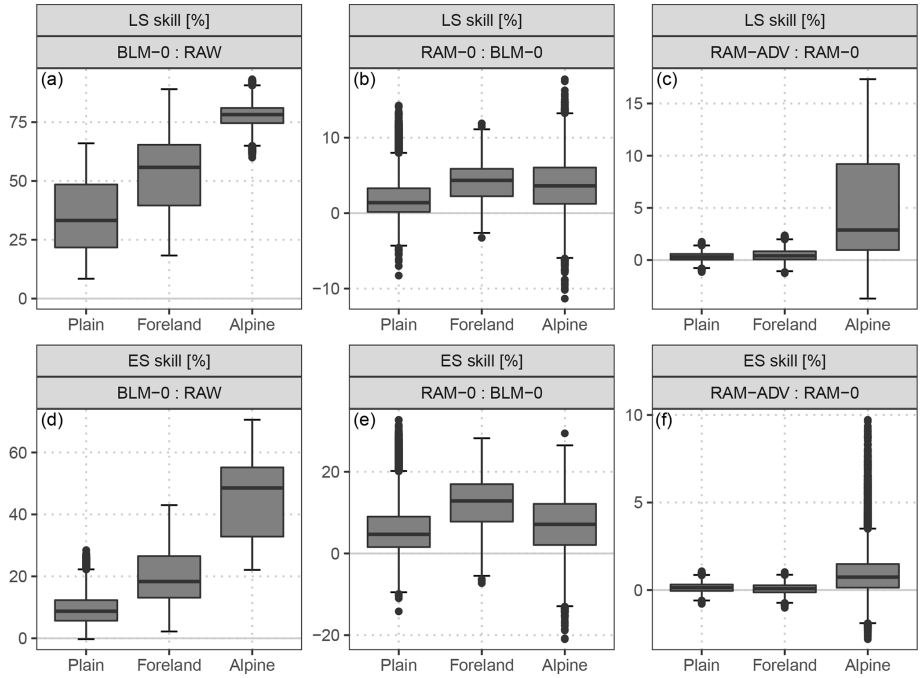

Figure 9. Aggregated skill scores (LS: **a–c**, ES: **d–f**) for groups of respective five weather stations which are located in the plain, in the mountain foreland near the Alps, or within an alpine valley. Each box-whisker contains boot-strapped mean values of the forecast steps from +12 to +72 h at a 12-hourly temporal resolution for all included stations. The scores are based on the full predictive bivariate distribution for the out-of-sample validation period. Compared are the BLM-0 model with the raw EPS as a reference, setup RAM-0 with setup BLM-0 as a reference, and the RAM-ADV correlation specification with the RAM-0 correlation model as a reference. Skill scores are in percent; positive values indicate improvements over the reference.

the two-dimensional location and scale parameters. This allows the statistical model to adjust for potential misspecifications in the ensemble wind direction by a smooth rotation conditional on the day of the year and the forecasted wind direction. For stations in complex terrain, this may be particularly advantageous due to unresolved topographical features.

The estimated effects confirm a distinct wind rotation for the valley site (Innsbruck), while for the station in the plain (Hamburg) barely any adjustments of the forecasted wind direction are needed (see Fig. 4). In terms of predictive performance, the more flexible model, RAM-0, outperforms the baseline model, BLM-0, for almost all times and stations (see Fig. 9b, e). However, the increase in predictive skill is similar for all three station types. This indicates that – even if no or only little rotation is needed – additional covariates usually yield a better adjustment of the distribution parameters and therefore an increased predictive skill. Furthermore, the results indicate that EPS wind forecasts in complex terrain are not solely tilted due to unresolved valley topographies, but show little skill on average. Thus, for alpine valley sites the rotation-allowing model mainly captures climatological properties conditional on the forecasted EPS wind direction. In accordance with this analysis, larger improvements can be found for stations located in the mountain foreland where the EPS has a higher information content and a certain rotation might be necessary.

These findings are supported by an additional comparison against the model inspired by Pinson (2012), which only uses a linear transformation of both wind components for the location parameters. The results show that a more flexible rotation-allowing specification is required to capture strong wind distortions; the full comparison is shown in Appendix C2.

5.2 Correlation specifications

Several different model specifications for the correlation parameter have been tested, among others a flexible setup employing wind direction and speed as potential covariates for the correlation parameter by nonlinear smooth effects following the idea of Schuhen et al. (2012). The estimated correlation parameters seem to be reasonable, and show, on average, larger values for Innsbruck than for Hamburg (see Fig. 6 for forecast step +12 h). In terms of predictive skill, all models tested show only minor improvements compared to the models with zero correlation. The improvements are highest for stations located inside alpine valleys, with a mean improvement of 1 % in terms of ES and 5 % in terms of LS skill scores (see Fig. 9c, f). The relatively small benefits due to the

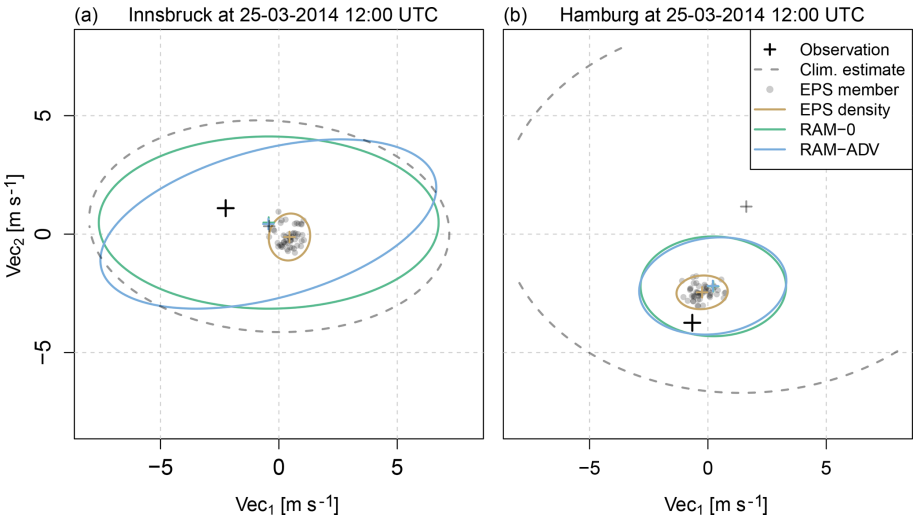

Figure 10. Two exemplary forecasts showing the respective observation (black cross), the climatological estimate (gray dashed line), the EPS member forecasts (gray points) and their empirical density (brown line), and the estimated bivariate distributions for the setups RAM-0 and RAM-ADV, without (green line) and with (blue line) modeled correlation, respectively. The climatological estimate uses the mean, the standard deviation, and the correlation of the observed wind components as bivariate distribution parameters. The lines show the 95 % percentiles of the respective bivariate distribution, the small crosses the ellipsoid centers or the location parameters. The shown observations and forecasts are valid at Innsbruck **(a)** and Hamburg **(b)** for 25 March 2014 12:00 UTC (forecast step +12 h). The forecasts are characteristic for **(a)** a valley station and **(b)** a station in the plain within our study.

explicit estimation of the correlation are confirmed by the results of an additionally tested model based on Schuhen et al. (2012) (Appendix C1).

As an illustration of the potential reasons for no more pronounced enhancements by an explicit estimation of the dependence structure, Fig. 10 shows exemplary forecasts for a station located within an alpine valley (Innsbruck; Fig. 10a) and in the plains (Hamburg; Fig. 10b). The figure shows the raw EPS members (gray points) plus the respective observations (black crosses), climatological estimates (gray dashed lines), and the corresponding post-processed bivariate distributions without (green lines) and with (blue lines) an explicitly estimated correlation parameter. For the valley station, the raw EPS has only little skill and the uncertainty of the post-processed bivariate distributions tends towards the climatological estimate. Although a distinct correlation is estimated by the RAM-ADV model, the variance is still in the same range as for the RAM-0 model. In contrast, for the station in the plain the uncertainty of the post-processed predictions is much smaller than the uncertainty of the climatological estimate due to a higher information content of the EPS. The estimated correlation is close to zero and the predictions of RAM-0 and RAM-ADV look almost identical with a similar elliptic shape as the raw EPS. This means that for locations where the ensemble provides only little information, the post-processed uncertainty is rather large and the statistical model tries to capture unexplained features by the correlation parameter. For stations where the predictive skill of the raw ensemble is already high, the statistical models get valuable information about the expected wind situation and are able to accurately specify the location and scale parameters. Thus, the correlation of the residuals becomes less important and typically smaller. This interpretation is supported by the probabilistic scores used in this study which show improvements in the RAM-ADV models mainly for alpine valley sites where the skill of the raw ensemble is rather low.

5.3 Proposed model specification

The study shows that the flexible rotation-allowing models bring significant performance benefits for stations located in complex terrain as well as for stations in the plain. Therefore, we propose using a similar setup employing both EPS wind components by a smooth rotation-allowing framework. For correlation, we have not found a clear distinction between the different correlation models tested for stations located in the plain and the foreland. For stations located within an alpine valley, minor improvements could be found. Despite these somewhat unexpected findings, this has clear advantages for operational usage: estimating a single bivariate response distribution forcing the correlation dependence structure to zero is the same as post-processing each wind component separately in a univariate setup with marginal Gaussian response distributions. A univariate post-processing approach for each respective wind component simplifies the estimation process in terms of complexity of the required statistical models and reduces computational time with only little loss of predictive skill, at least for the stations tested in this study.

Code availability. The bivariate Gaussian model estimation is performed in R 3.5.2 (R Core Team, 2018) based on the R package bamlss (Umlauf et al., 2018). The package provides a flexible toolbox for distribution regression models in a Bayesian framework. Introductory material and example code on how to set up the models as presented in this article can be found at http://bayesr.r-forge.r-project.org (last access: 28 June 2019). The computation of the ES is based on the R package scoringRules (Jordan et al., 2019).

Appendix A: Model specification complements

A1 Smooth functions

Generalized additive models (GAMs, Hastie and Tibshirani, 1986) and generalized additive models for location, shape, and scale (GAMLSS, Rigby and Stasinopoulos, 2005) are generalizations of linear regression models which allow one to include potentially nonlinear (and even multidimensional) effects in the linear predictors η. Nonlinear terms are frequently approximated by smooth functions, also referred to as regression splines. These regression splines are directly linked to the model parameters as additive terms in η and allow the statistical model to include nonlinear transformations of a specific covariate, if needed. For further details a comprehensive introduction to GAMs is given in Wood (2017). An example of an additive predictor η with a smooth function is

$$\eta = \alpha_0 + \underbrace{f_1(x_2)\alpha_1 \cdot x_1}_{\text{linear effect for } x_1} + \underbrace{f_1(x_2)}_{\text{pot. nonlinear effect for } x_2}, \tag{A1}$$

where $\alpha_\bullet$ are regression coefficients, $x_\bullet$ the covariates, and $\alpha_1 \cdot x_1$ and $f_1(x_2)$ a linear and a potentially nonlinear one-dimensional effect, respectively. Generally, f_1 can be any transformation of the covariate x_2 dependent on the specification of f_1. For periodic values smooth "cyclic" splines are often applied, meaning that the function has the same value at its upper and lower boundaries. This is similar to applying a linear combination of (several) trigonometric functions, as, e.g., performed by Schuhen et al. (2012). In this study, we utilize "cubic" smooth functions with cyclic constraints. A detailed description of cyclic cubic regression splines is given in Wood (2017, chap. 4.1.3).

A2 Time-adaptive training scheme

To account for seasonal variations of the intercept and the linear coefficients, seasonal cyclic splines are used. If the covariates provide sufficient information, a time-adaptive training scheme might not be required. However, if the bias and/or the slope coefficient are not constant throughout the year or the covariate's skill varies over the year, these terms are mandatory to allow the statistical model to depict seasonal features.

We therefore fit one statistical model over a training data set including several years of data, but allow the coefficient included in the linear predictor(s) η to smoothly evolve over the year:

$$\eta = \underbrace{\alpha_0 + f_{\star 0}(\text{doy})}_{\substack{\text{seasonally varying} \\ \text{intercept}}} + \underbrace{(\alpha_1 + f_1(\text{doy}))}_{\substack{\text{seasonally varying} \\ \text{coefficient for } x_1}} \cdot x_1 + \cdots + \underbrace{(\alpha_n + f_n(\text{doy}))}_{\substack{\text{seasonally varying} \\ \text{coefficient for } x_n}} \cdot x_n. \tag{A2}$$

As before, $\alpha_\bullet$ are the regression coefficients, $x_\bullet$ are the covariates, and $f_\bullet(\text{doy})$ employ cyclic regression splines conditional on the day of the year (doy). Within this study, we refer to the regression coefficients $\alpha_\bullet$ also as global intercept or slope coefficients to emphasize that they are unconditional on the day of the year.

Appendix B: Skill scores used for verification

To compare the different bivariate Gaussian models of this study, we employ skill scores. A skill score shows the improvements over a reference. For all measures with a perfect score of zero, the skill score simplifies to

$$\text{skill score} = \frac{\text{score}_{\text{fcst}} - \text{score}_{\text{ref}}}{\text{score}_{\text{opt}} - \text{score}_{\text{ref}}} = 1 - \frac{\text{score}_{\text{fcst}}}{\text{score}_{\text{ref}}}, \tag{B1}$$

where $\text{score}_{\text{fcst}}$ is the forecast's score, $\text{score}_{\text{opt}} = 0$ refers to a hypothetical optimal or perfect score, and $\text{score}_{\text{ref}}$ is the score for the reference (Gneiting and Raftery, 2007).

In this study we use the logarithmic score (LS, Good, 1952) and the energy score (ES, Gneiting and Raftery, 2007) to validate the probabilistic performance of the bivariate Gaussian predictions of the statistical post-processing models. Both multivariate scores evaluate the full predictive distribution returned by the statistical models.

The calculation of the ES is based on the R package **scoringRules** (Jordan et al., 2019). For a predictive distribution f on $\mathbb{R}^d$ given through m discrete samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ from f with $\boldsymbol{X}_i = (X_i^{(1)}, \ldots, X_i^{(d)}) \in \mathbb{R}^d, i = 1, \ldots, m$, the ES can be written as

$$\text{ES}(f, \boldsymbol{y}) = \frac{1}{m}\sum_{i=1}^{m} \|\boldsymbol{X}_i - \boldsymbol{y}\| - \frac{1}{2m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} \|\boldsymbol{X}_i - \boldsymbol{X}_j\|, \tag{B2}$$

where $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^d$ and $\boldsymbol{y} = (y^{(1)}, \ldots, y^{(d)}) \in \mathbb{R}^d$ the multivariate observation. The calculation of the ES for all post-processed forecasts is based on $m = 1000$ random draws from the bivariate Gaussian distribution.

The logarithmic score is defined based on the log-density (or log-likelihood):

$$\text{LS}(f, \boldsymbol{y}) = \log(f(\boldsymbol{y})), \tag{B3}$$

where the probabilistic forecast on $\mathbb{R}^d$ is given by the probability density function f, and $\boldsymbol{y} = (y^{(1)}, \ldots, y^{(d)}) \in \mathbb{R}^d$ is the multivariate observation (Jordan et al., 2019). However, it is not possible to compute skill scores based on the LS from Eq. (B3). The reason is that the LS for a perfect prediction is not zero but actually diverges to infinity. To mitigate this shortcoming we follow Lindsey (1996, chap. 3.2.1) and define the likelihood based on the cumulative distribution function F on an interval of length $\boldsymbol{\Delta}$ rather than the density at a single point:

$$f(\boldsymbol{y}) \approx \frac{F(\boldsymbol{y} + \boldsymbol{\Delta}/2) - F(\boldsymbol{y} - \boldsymbol{\Delta}/2)}{\boldsymbol{\Delta}}, \tag{B4}$$

where $\boldsymbol{\Delta}$ is set to the precision of reporting $\boldsymbol{y}$. Employing this idea, the LS can be approximated (up to a constant) by

$$\mathrm{LS}(f, \boldsymbol{y}) \propto \log(F(\boldsymbol{y} + \boldsymbol{\Delta}/2) - F(\boldsymbol{y} - \boldsymbol{\Delta}/2)). \tag{B5}$$

This representation has the advantage that for a perfect fit the LS is zero as $F(\boldsymbol{y} + \boldsymbol{\Delta}/2) = 1$ and $F(\boldsymbol{y} - \boldsymbol{\Delta}/2) = 0$ so that $\mathrm{LS} = \log(1 - 0) = 0$. Consequently, the corresponding skill scores can be computed as shown in Eq. (B1). In this study, we use $\boldsymbol{\Delta}/2 = 0.1\,\mathrm{m\,s^{-1}}$; the computation of the bivariate cumulative distribution function is based on the R package **mvtnorm** (Genz and Bretz, 2009). For the discrete raw ensemble forecasts of wind vectors, the empirical mean and covariance matrix of the ensemble are used to calculate the LS as in Eq. (B5). This implicitly assumes that the raw ensemble wind forecasts follow a bivariate Gaussian distribution.

Appendix C: Further model comparisons

To benchmark the models as presented in this study, we compare our specifications to those of Schuhen et al. (2012) and Pinson (2012).

C1 Comparison with Schuhen et al. (2012)

Schuhen et al. (2012) fit a bivariate Gaussian model for the wind components in three phases. First, they fit the correlation parameter as a trigonometric function of the ensemble mean wind direction by weighted nonlinear least squares. They estimate the regression coefficients for the correlation parameter offline in a pre-processing step for a separate year, either for a single site or a group of stations. The adjustment for a suitable number of trigonometric cycles must be done manually, which can be prone to errors according to Schuhen et al. (2012, p. 3207). Second, univariate models are fitted for the components of the two-dimensional location parameter by standard linear regression. Third, the two-dimensional variance parameter of the bivariate Gaussian distribution is estimated by maximum likelihood keeping all other parameters fixed. In contrast to the first phase, the estimation within the second and third phases is performed online using a rolling training period, either for a single site or a group of stations.

In this study, we apply Schuhen et al. (2012) using the implementation of Lerch (2019). As the focus of the current paper is on post-processing wind vector forecasts for stations with different site characteristics, we perform the estimation for each station separately. We use a rolling training period of 40 days and employ two periods for the trigonometric function in the estimation of the correlation parameter on the training data set. Figure C1 shows the comparison to the baseline model BLM-0, to the rotation-allowing model without correlation RAM-0, and to the rotation-allowing model with correlation RAM-ADV for the out-of-sample validation period of this study. The predictive performance of Schuhen et al. (2012) is overall comparable to the BLM-0 setup. Accordingly, the comparison to RAM-0 and RAM-ADV confirms that within this study the largest potential for improvement lies in the correct specification of the location and scale parameters of the bivariate Gaussian distribution.

C2 Comparison with the Pinson (2012)-type model

Since one of the major aspects within this study is the rotation of the wind direction, we compare our models to a model inspired by Pinson (2012), which also uses both wind components of the raw ensemble as predictors for both components of the bivariate location parameter. We define the two-dimensional location and scale part according to Pinson (2012), but employ our model framework and fix the correlation to zero, i.e., by

$$\begin{aligned}
\mu_\star &= \alpha_{\star 0} + f_{\star 0}(\mathrm{doy}) + (\alpha_{\star 1} + f_{\star 1}(\mathrm{doy})) \cdot \mathrm{vec}_{1,\mathrm{mean}} \\
&\quad + (\alpha_{\star 2} + f_{\star 2}(\mathrm{doy})) \cdot \mathrm{vec}_{2,\mathrm{mean}}, \\
\log(\sigma_\star) &= \beta_{\star 0} + g_{\star 0}(\mathrm{doy}) + (\beta_{\star 1} + g_{\star 1}(\mathrm{doy})) \\
&\quad \cdot \mathrm{vec}_{\star,\mathrm{log.sd}}, \\
\mathrm{rhogit}(\rho) &= 0,
\end{aligned} \tag{C1}$$

where, as before, $\alpha_\bullet$ and $\beta_\bullet$ are regression coefficients and $f_\bullet$ and $g_\bullet$ are cyclic regression splines. The location part employs a linear transformation of the wind components, which is able to rotate the wind direction but in a restricted linear manner.

Figure C2 shows the comparison of the Pinson-type model to this study's baseline model BLM-0, to the rotation-allowing model without correlation RAM-0, and to the rotation-allowing model with correlation RAM-ADV for the out-of-sample validation period. The results show that the Pinson-type model is apparently a mixture of the BLM-0 and RAM-0 models. Hence, for minor distortions in wind directions, such as in the foreland, the Pinson-type model is already sufficient and has clear benefits when compared to the non-rotation-allowing BLM-0 model. For stations in complex terrain, the RAM-0 model shows clear advantages over the less flexible Pinson-type model. This indicates that a more flexible rotation-allowing specification is required to capture strong wind distortions, e.g., due to discrepancies between the model and real topography. The explicit estimation of the correlation (RAM-ADV) further increases the performance, but mainly for alpine stations (see Sect. 5.2).

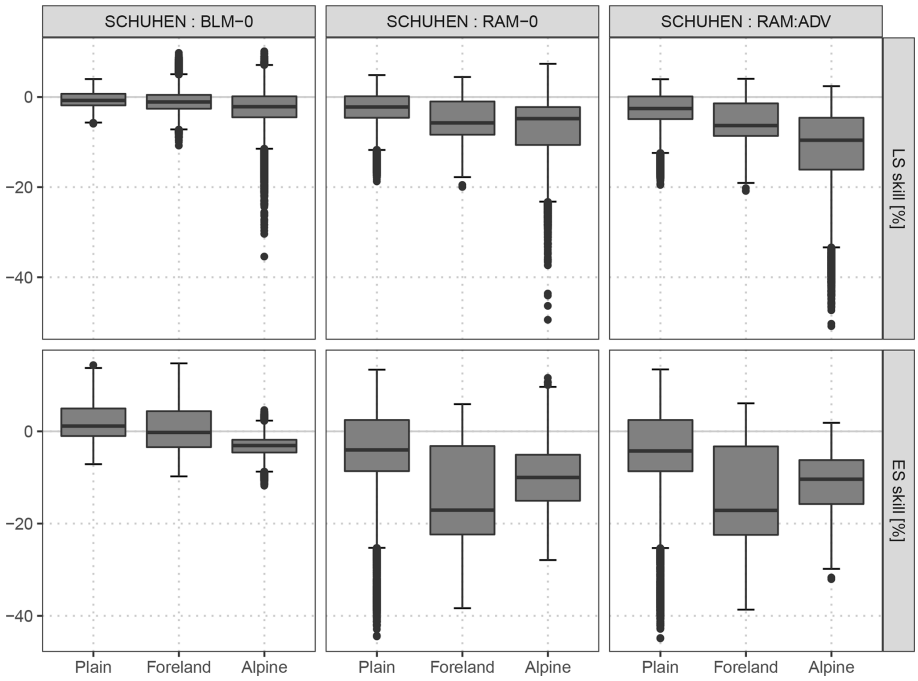

Figure C1. As Fig. 9, but an adaption of the Schuhen et al. (2012) model is compared to the baseline model BLM-0, to the rotation-allowing model without correlation RAM-0, and to the rotation-allowing model with correlation RAM-ADV.

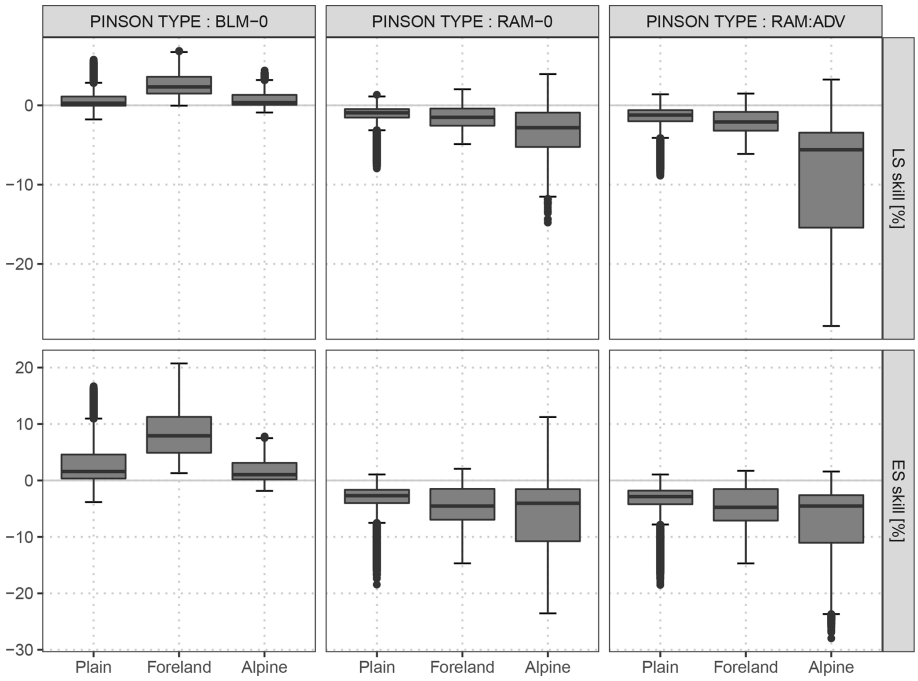

Figure C2. As Fig. 9, but the model specification inspired by Pinson (2012) (Eq. C1) is compared to the baseline model BLM-0, to the rotation-allowing model without correlation RAM-0, and to the rotation-allowing model with correlation RAM-ADV.

Author contributions. This study is based on the PhD work of MNL under supervision of GJM and AZ. The majority of the work for this study was performed by MNL with strong guidance of RS. All the authors worked closely together in discussing the results and commented on the paper.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This project was funded by the Austrian Research Promotion Agency (FFG), grant no. 858537. We also thank the Zentralanstalt für Meteorologie und Geodynamik (ZAMG) for providing access to the data. Furthermore, we are grateful to the editor and the reviewers for their valuable comments.

Financial support. This research has been supported by the Austrian Research Promotion Agency (FFG) (grant no. 858537).

Review statement. This paper was edited by Christopher Paciorek and reviewed by Sebastian Lerch and one anonymous referee.

References

Baran, S.: Probabilistic Wind Speed Forecasting Using Bayesian Model Averaging with Truncated Normal Components, Comput. Stat. Data An., 75, 227–238, https://doi.org/10.1016/j.csda.2014.02.013, 2014.

Baran, S. and Lerch, S.: Log-Normal Distribution Based Ensemble Model Output Statistics Models for Probabilistic Wind-Speed Forecasting, Q. J. Roy. Meteor. Soc., 141, 2289–2299, https://doi.org/10.1002/qj.2521, 2015.

Baran, S. and Lerch, S.: Mixture EMOS Model for Calibrating Ensemble Forecasts of Wind Speed, Environmetrics, 27, 116–130, https://doi.org/10.1002/env.2380, 2016.

Buizza, R., Houtekamer, P. L., Pellerin, G., Toth, Z., Zhu, Y., and Wei, M.: A Comparison of the ECMWF, MSC, and NCEP Global Ensemble Prediction Systems, Mon. Weather Rev., 133, 1076–1097, https://doi.org/10.1175/MWR2905.1, 2005.

Courtney, J. F., Lynch, P., and Sweeney, C.: High Resolution Forecasting for Wind Energy Applications Using Bayesian Model Averaging, Tellus A, 65, 19669, https://doi.org/10.3402/tellusa.v65i0.19669, 2013.

Eide, S. S., Bremnes, J. B., and Steinsland, I.: Bayesian Model Averaging for Wind Speed Ensemble Forecasts Using Wind Speed and Direction, Weather Forecast., 32, 2217–2227, https://doi.org/10.1175/WAF-D-17-0091.1, 2017.

EuropeanCommission: Time Based Separation at Heathrow, available at: https://ec.europa.eu/transport/modes/air/ses/ses-award-2016/projects/time-based-separation-heathrow_en, (last access: 16 February 2019), 2018.

Gamerman, D.: Sampling from the Posterior Distribution in Generalized Linear Mixed Models, Stat. Comput., 7, 57–68, https://doi.org/10.1023/A:1018509429360, 1997.

Gebetsberger, M., Messner, J. W., Mayr, G. J., and Zeileis, A.: Fine-Tuning Nonhomogeneous Regression for Probabilistic Precipitation Forecasts: Unanimous Predictions, Heavy Tails, and Link Functions, Mon. Weather Rev., 145, 4693–4708, https://doi.org/10.1175/MWR-D-16-0388.1, 2017.

Genz, A. and Bretz, F.: Computation of Multivariate Normal and t Probabilities, Lecture Notes in Statistics, Springer-Verlag, Heidelberg, Germany, 2009.

Glahn, H. R. and Lowry, D. A.: The Use of Model Output Statistics (MOS) in Objective Weather Forecasting, J. Appl. Meteorol., 11, 1203–1211, https://doi.org/10.1175/1520-0450(1972)011<1203:TUOMOS>2.0.CO;2, 1972.

Gneiting, T.: Editorial: Probabilistic Forecasting, J. R. Stat. Soc. A Stat., 171, 319–321, https://doi.org/10.1111/j.1467-985X.2007.00522.x, 2008.

Gneiting, T. and Katzfuss, M.: Probabilistic Forecasting, Ann. Rev. Stat. Appl., 1, 125–151, https://doi.org/10.1146/annurev-statistics-062713-085831, 2014.

Gneiting, T. and Raftery, A. E.: Strictly Proper Scoring Rules, Prediction, and Estimation, J. Am. Stat. Assoc., 102, 359–378, https://doi.org/10.1198/016214506000001437, 2007.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L., and Johnson, N. A.: Assessing Probabilistic Forecasts of Multivariate Quantities, with an Application to Ensemble Predictions of Surface Winds, TEST, 17, 211–235, https://doi.org/10.1007/s11749-008-0114-x, 2008.

Good, I. J.: Rational Decisions, J. Roy. Stat. Soc. B Met., 14, 107–114, 1952.

Hastie, T. and Tibshirani, R.: Generalized Additive Models, Stat. Sci., 1, 297–310, 1986.

Jordan, A., Krüger, F., and Lerch, S.: Evaluating Probabilistic Forecasts with scoringRules, J. Stat. Softw., accepted, 2019.

Klein, N., Kneib, T., Klasen, S., and Lang, S.: Bayesian Structured Additive Distributional Regression for Multivariate Responses, J. R. Stat. Soc. C-Appl., 64, 569–591, https://doi.org/10.1111/rssc.12090, 2014.

Kunkel, K. E., Karl, T. R., Brooks, H., Kossin, J., Lawrimore, J. H., Arndt, D., Bosart, L., Changnon, D., Cutter, S. L., Doesken, N., Emanuel, K., Groisman, P. Y., Katz, R. W., Knutson, T., O'Brien, J., Paciorek, C. J., Peterson, T. C., Redmond, K., Robinson, D., Trapp, J., Vose, R., Weaver, S., Wehner, M., Wolter, K., and Wuebbles, D.: Monitoring and Understanding Trends in Extreme Storms: State of Knowledge, B. Am. Meteorol. Soc., 94, 499–514, https://doi.org/10.1175/BAMS-D-11-00262.1, 2012.

Lerch, S.: Bivariate EMOS Model for Wind Vectors of Schuhen et al. (2012), available at: https://github.com/slerch/bivariate_EMOS, last access: 16 May 2019.

Lerch, S. and Thorarinsdottir, T. L.: Comparison of Non-Homogeneous Regression Models for Probabilistic Wind Speed Forecasting, Tellus A, 65, 21206, https://doi.org/10.3402/tellusa.v65i0.21206, 2013.

Lindsey, J. K.: Parametric Statistical Inference, Oxford University Press, Oxford, New York, USA, 1996.

Messner, J. W., Mayr, G. J., Wilks, D. S., and Zeileis, A.: Extending Extended Logistic Regression: Extended versus Separate versus Ordered versus Censored, Mon. Weather Rev., 142, 3003–3014, https://doi.org/10.1175/MWR-D-13-00355.1, 2014a.

Messner, J. W., Mayr, G. J., Zeileis, A., and Wilks, D. S.: Heteroscedastic Extended Logistic Regression for Postprocessing

of Ensemble Guidance, Mon. Weather Rev., 142, 448–456, https://doi.org/10.1175/mwr-d-13-00271.1, 2014b.

NASA JPL: NASA Shuttle Radar Topography Mission Global 30 Arc Second [Data Set], NASA EOSDIS Land Processes DAAC, https://doi.org/10.5067/MEaSUREs/SRTM/SRTMGL30.002, 2013.

Palmer, T. N.: The Economic Value of Ensemble Forecasts as a Tool for Risk Assessment: From Days to Decades, Q. J. Roy. Meteor. Soc., 128, 747–774, https://doi.org/10.1256/0035900021643593, 2002.

Pinson, P.: Adaptive Calibration of (u,v)-Wind Ensemble Forecasts, Q. J. Roy. Meteor. Soc., 138, 1273–1284, https://doi.org/10.1002/qj.1873, 2012.

Pinson, P. and Tastu, J.: Discrimination Ability of the Energy Score, Report, Technical University of Denmark (DTU), Kgs. Lyngby, available at: http://orbit.dtu.dk/files/56966842/tr13_15_Pinson_Tastu.pdf (last access: 16 February 2019), 2013.

R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org, last access: 20 December 2018.

Rigby, R. A. and Stasinopoulos, D. M.: Generalized Additive Models for Location, Scale and Shape, J. R. Stat. Soc. C-Appl., 54, 507–554, https://doi.org/10.1111/j.1467-9876.2005.00510.x, 2005.

Schefzik, R., Thorarinsdottir, T. L., and Gneiting, T.: Uncertainty Quantification in Complex Simulation Models Using Ensemble Copula Coupling, Stat. Sci., 28, 616–640, https://doi.org/10.1214/13-STS443, 2013.

Scheuerer, M. and Möller, D.: Probabilistic Wind Speed Forecasting on a Grid Based on Ensemble Model Output Statistics, Ann. Appl. Stat., 9, 1328–1349, https://doi.org/10.1214/15-AOAS843, 2015.

Schuhen, N., Thorarinsdottir, T. L., and Gneiting, T.: Ensemble Model Output Statistics for Wind Vectors, Mon. Weather Rev., 140, 3204–3219, https://doi.org/10.1175/MWR-D-12-00028.1, 2012.

Sloughter, J. M., Gneiting, T., and Raftery, A. E.: Probabilistic Wind Speed Forecasting Using Ensembles and Bayesian Model Averaging, J. Am. Stat. Assoc., 105, 25–35, https://doi.org/10.1198/jasa.2009.ap08615, 2010.

Taillardat, M., Mestre, O., Zamo, M., and Naveau, P.: Calibrated Ensemble Forecasts Using Quantile Regression Forests and Ensemble Model Output Statistics, Mon. Weather Rev., 144, 2375–2393, https://doi.org/10.1175/MWR-D-15-0260.1, 2016.

Thorarinsdottir, T. L. and Gneiting, T.: Probabilistic Forecasts of Wind Speed: Ensemble Model Output Statistics by Using Heteroscedastic Censored Regression, J. R. Stat. Soc. A Stat., 173, 371–388, https://doi.org/10.1111/j.1467-985X.2009.00616.x, 2010.

Umlauf, N., Klein, N., and Zeileis, A.: BAMLSS: Bayesian Additive Models for Location, Scale, and Shape (and Beyond), J. Comput. Graph. Stat., 27, 612–627, https://doi.org/10.1080/10618600.2017.1407325, 2018.

Vislocky, R. L. and Fritsch, J. M.: Generalized Additive Models versus Linear Regression in Generating Probabilistic MOS Forecasts of Aviation Weather Parameters, Weather Forecast., 10, 669–680, https://doi.org/10.1175/1520-0434(1995)010<0669:GAMVLR>2.0.CO;2, 1995.

Vose, R. S., Applequist, S., Bourassa, M. A., Pryor, S. C., Barthelmie, R. J., Blanton, B., Bromirski, P. D., Brooks, H. E., DeGaetano, A. T., Dole, R. M., Easterling, D. R., Jensen, R. E., Karl, T. R., Katz, R. W., Klink, K., Kruk, M. C., Kunkel, K. E., MacCracken, M. C., Peterson, T. C., Shein, K., Thomas, B. R., Walsh, J. E., Wang, X. L., Wehner, M. F., Wuebbles, D. J., and Young, R. S.: Monitoring and Understanding Changes in Extremes: Extratropical Storms, Winds, and Waves, B. Am. Meteorol. Soc., 95, 377–386, https://doi.org/10.1175/BAMS-D-12-00162.1, 2013.

WindEurope: Wind Energy in Europe Scenarios for 2030, Tech. rep., available at: https://windeurope.org/about-wind/reports/wind-energy-in-europe-scenarios-for-2030/ (last access: 16 February 2019), 2017.

Wood, S. N.: Generalized Additive Models: An Introduction with R, Chapman and Hall/CRC, https://doi.org/10.1201/9781315370279, 2017.

Article VIII

Zeileis A., Fisher J.C., Hornik K., Ihaka R., McWhite C.D., Murrell P., Stauffer R., and Wilke C.O. (2020). *colorspace: A Toolbox for Manipulating and Assessing Colors and Palettes*. Journal of Statistical Software, 96(1), 1–49, doi:10.18637/jss.v096.i01.

JCR ranking: **Category 1** in *Statistics and Probability*.

Contribution (CRT): *Conceptualization / software / visualization / writing, review and editing*.

Journal of Statistical Software

November 2020, Volume 96, Issue 1. doi: 10.18637/jss.v096.i01

colorspace: A Toolbox for Manipulating and Assessing Colors and Palettes

Achim Zeileis
Universität Innsbruck

Jason C. Fisher
U.S. Geological Survey

Kurt Hornik
WU Wirtschaftsuniversität Wien

Ross Ihaka
University of Auckland

Claire D. McWhite
The University of Texas at Austin

Paul Murrell
University of Auckland

Reto Stauffer
Universität Innsbruck

Claus O. Wilke
The University of Texas at Austin

Abstract

The R package **colorspace** provides a flexible toolbox for selecting individual colors or color palettes, manipulating these colors, and employing them in statistical graphics and data visualizations. In particular, the package provides a broad range of color palettes based on the HCL (hue-chroma-luminance) color space. The three HCL dimensions have been shown to match those of the human visual system very well, thus facilitating intuitive selection of color palettes through trajectories in this space. Using the HCL color model, general strategies for three types of palettes are implemented: (1) Qualitative for coding categorical information, i.e., where no particular ordering of categories is available. (2) Sequential for coding ordered/numeric information, i.e., going from high to low (or vice versa). (3) Diverging for coding ordered/numeric information around a central neutral value, i.e., where colors diverge from neutral to two extremes. To aid selection and application of these palettes, the package also contains scales for use with **ggplot2**, **shiny** and **tcltk** apps for interactive exploration, visualizations of palette properties, accompanying manipulation utilities (like desaturation and lighten/darken), and emulation of color vision deficiencies. The **shiny** apps are also hosted online at http://hclwizard.org/.

Keywords: color, palette, HCL, RGB, hue, color vision deficiency, R.

1. Introduction

Color is an integral element of many statistical graphics and data visualizations. Therefore, colors should be carefully chosen to support all viewers in accessing the information displayed

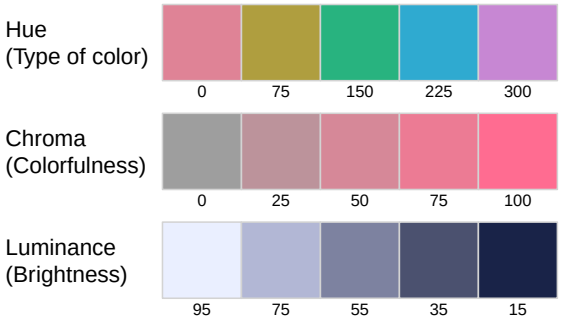

Figure 1: Axes of the HCL color space. Top: Hue H changes from 0 (red) via 75 (yellow), etc. to 300 (purple) with fixed $C = 60$ and $L = 65$. Center: Chroma C changes from 0 (gray) to 100 (colorful) with fixed $H = 0$ (red) and $L = 65$. Bottom: Luminance L changes from 95 (light) to 15 (dark) with fixed $H = 260$ (blue) and $C = 25$ (low, close to gray).

(Tufte 1990; Brewer 1999; Ware 2004; Wilkinson 2005; Wilke 2019). However, until relatively recently many software packages have been using color palettes derived from simple RGB (red-green-blue) color combinations such as the RGB "rainbow" (or "jet") color palette with poor perceptual properties. See Hawkins, McNeall, Stephenson, Williams, and Carlson (2014) and Stauffer, Mayr, Dabernig, and Zeileis (2015) and the references therein for an overview.

To address these problems, many improved color palettes with better perceptual properties have been receiving increasing attention in the literature (Harrower and Brewer 2003; Zeileis, Hornik, and Murrell 2009; Smith and Van der Walt 2015; CARTO 2019; Crameri 2018). Many systems for statistical and scientific computing provide infrastructure for such color palettes. For example, for R (R Core Team 2020) the list of useful packages encompasses **RColorBrewer** (Neuwirth 2014), **viridis** (Garnier 2018), **rcartocolor** (Nowosad 2019), **wesanderson** (Ram and Wickham 2018), and **scico** (Pedersen and Crameri 2020) among many others. Furthermore, packages like **pals** (Wright 2019) and **paletteer** (Hvitfeldt 2020) collect many of the proposed palettes in combination with a unified interface. Most of these palettes, however, are pre-existing palettes, stored as a limited set of colors and interpolated as necessary. And even if specific algorithms have been used in the initial construction of the palettes, these are often not reflected in the software implementations.

The **colorspace** package (Ihaka *et al.* 2020) adopts a somewhat different approach that gives the user direct access to the construction principles underlying its palettes. These are based on simple trajectories in the perceptually-based HCL (hue-chroma-luminance) color space (Wikipedia 2020e) whose axes match those of the human visual system very well: Hue (type of color, dominant wavelength), chroma (colorfulness), luminance (brightness), see Figure 1. Thus, utilizing this color model the **colorspace** package can derive general and adaptable strategies for color palettes; manipulate individual colors and color palettes; and assess and visualize the properties of color palettes (beyond simple color swatches). Specifically, **colorspace** provides three types of palettes based on the HCL model:

- *Qualitative:* Designed for coding categorical information, i.e., where no particular ordering of categories is available and every color should receive the same perceptual weight. Function: `qualitative_hcl()`.
- *Sequential:* Designed for coding ordered/numeric information, i.e., where colors go from high to low (or vice versa). Function: `sequential_hcl()`.

- *Diverging:* Designed for coding ordered/numeric information around a central neutral value, i.e., where colors diverge from neutral to two extremes. Function: `diverging_hcl()`.

A broad collection of prespecified palettes is shipped in the package. In addition, existing palettes can be easily tweaked and new or adapted palettes registered. The prespecified palettes include suitable HCL color choices that closely approximate most palettes from packages **RColorBrewer**, **rcartocolor**, and **viridis** by using only a small set of hue, chroma, and luminance parameters.

To aid choice and application of these palettes the package provides (a) scales for use with **ggplot2** (Wickham 2016), (b) **shiny** (Chang, Cheng, Allaire, Xie, and McPherson 2020) and **tcltk** (R Core Team 2020) apps for interactive exploration, (c) visualizations of palette properties, and (d) accompanying manipulation utilities (like converting to grayscale by desaturation, lighten/darken, and emulation of color vision deficiencies).

The remainder of the paper is organized as follows: Section 2 gives a first overview of the package's "look & feel" and the general workflow. Section 3 summarizes the S4 color space classes and methods in the package. Section 4 introduces the extensible collection of HCL-based palettes along with their construction details. Section 5 presents the toolbox for palette visualization and assessment. Section 6 discusses the implemented techniques for color vision deficiency emulation that help assess the suitability of colors for colorblind viewers. Section 7 briefly highlights the interactive color apps from the package. Some further color manipulation utilities are highlighted in Section 8 before Section 9 concludes the paper.

2. A quick tour

The stable release version of **colorspace** is hosted on the Comprehensive R Archive Network (CRAN) at https://CRAN.R-project.org/package=colorspace and the development version is hosted on R-Forge at https://R-Forge.R-project.org/projects/colorspace/.

2.1. Choosing HCL-based color palettes

The **colorspace** package ships with a wide range of predefined color palettes, specified through suitable trajectories in the HCL (hue-chroma-luminance) color space. A quick overview can be gained easily with the `hcl_palettes()` function (see Figure 2, some of these are illustrated in more detail later):

```
R> library("colorspace")
R> hcl_palettes(plot = TRUE)
```

A suitable vector of colors can be easily computed by specifying the desired number of colors and the palette name (see Figure 2 for possible palette names), e.g.,

```
R> q4 <- qualitative_hcl(4, palette = "Dark 3")
R> q4

[1] "#E16A86" "#909800" "#00AD9A" "#9183E6"
```

Figure 2: Brief overview of available predefined palettes in **colorspace**.

The functions `sequential_hcl()`, and `diverging_hcl()` work analogously. Additionally, a palette's hue/chroma/luminance parameters can be modified, thus allowing for easy customization of each palette. Moreover, the `choose_palette()`/`hclwizard()` app provides convenient user interfaces to perform palette customization interactively. Finally, even more flexible diverging HCL palettes are provided by `divergingx_hcl()`.

2.2. Usage with base graphics

The color vectors returned by the HCL palette functions can usually be passed directly to most base graphics, typically through the `col` argument. Here, the `q4` vector created above is used in a time series display (see the left panel of Figure 3):

```
R> plot(log(EuStockMarkets), plot.type = "single", col = q4, lwd = 2)
R> legend("topleft", colnames(EuStockMarkets), col = q4, lwd = 3, bty = "n")
```

As another example for a sequential palette, we demonstrate how to create a spine plot (see the right panel of Figure 3) displaying the proportion of Titanic passengers that survived per class. The `"Purples 3"` palette is used, which is quite similar to the **ColorBrewer.org** (Harrower and Brewer 2003) palette `"Purples"`. Here, only two colors are employed: a dark purple that is highlighted against a light gray.

```
R> ttnc <- margin.table(Titanic, c(1, 4))
R> spineplot(ttnc, col = sequential_hcl(2, palette = "Purples 3"))
```

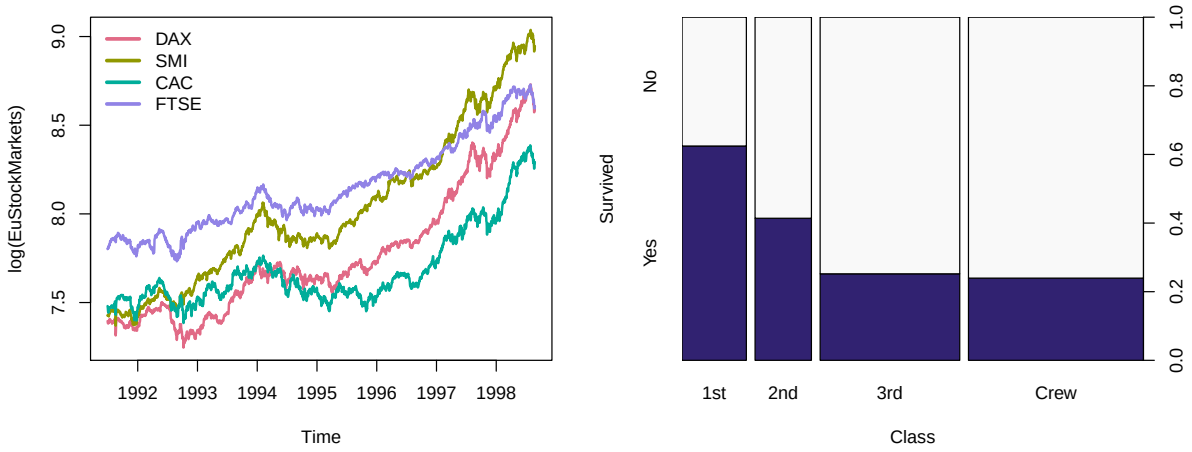

Figure 3: Using **colorspace** with base R graphics. Left: Time series plot of log-prices from `EuStockMarkets` data with `qualitative_hcl(4, "Dark 3")` palette. Right: Spine plot with survival proportions across passenger classes in the `titanic` data with `sequential_hcl(2, "Purples 3")` palette.

2.3. Usage with ggplot2

To provide access to the HCL color palettes from within **ggplot2** graphics (Wickham 2016; Wickham *et al.* 2020) suitable discrete, continuous, and binned **ggplot2** color scales are provided. The scales are named via the scheme

```
scale_<aesthetic>_<datatype>_<colorscale>()
```

where

- `<aesthetic>` is the name of the aesthetic (`fill`, `color`, `colour`).
- `<datatype>` is the type of the variable plotted (`discrete`, `continuous`, `binned`).
- `<colorscale>` sets the type of the color scale used (i.e., `qualitative`, `sequential`, `diverging`, `divergingx`).

To illustrate their usage two simple examples are shown using the qualitative `"Dark 3"` and sequential `"Purples 3"` palettes that were also employed above. For the first example, semi-transparent shaded densities of the sepal length from the `iris` data are shown, grouped by species (see the left panel of Figure 4).

```
R> library("ggplot2")
R> ggplot(iris, aes(x = Sepal.Length, fill = Species)) +
+    geom_density(alpha = 0.6) +
+    scale_fill_discrete_qualitative(palette = "Dark 3")
```

And for the second example the sequential palette is used to code the cut levels in a scatter of price by carat in the `diamonds` data (or rather a small subsample thereof, see the right panel of Figure 4). The scale function first generates six colors but then drops the first color because the light gray is too light here. (Alternatively, the chroma and luminance parameters could also be tweaked.)

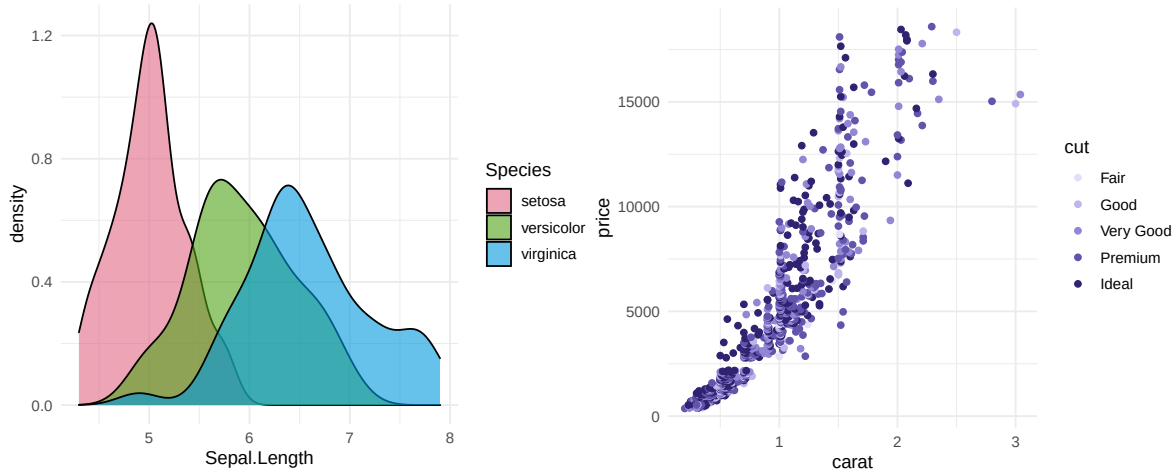

Figure 4: Using **colorspace** with **ggplot2** graphics. Left: Kernel density of sepal length, grouped and shaded by species, in the `iris` data with semi-transparent `scale_fill_discrete_qualitative(palette = "Dark 3")` color scale. Right: Scatter plot of price by carat, shaded by cut levels, in a subsample of the `diamonds` data with the `scale_color_discrete_sequential(palette = "Purples 3", nmax = 6, order = 2:6)` color scale.

```
R> dsamp <- diamonds[1 + 1:1000 * 50, ]
R> ggplot(dsamp, aes(carat, price, color = cut)) + geom_point() +
+    scale_color_discrete_sequential(palette = "Purples 3", nmax = 6,
+      order = 2:6)
```

2.4. Palette visualization and assessment

The **colorspace** package also provides a number of functions that aid visualization and assessment of its palettes.

- `demoplot()` can display a palette (with arbitrary number of colors) in a range of typical and somewhat simplified statistical graphics.
- `hclplot()` converts the colors of a palette to the corresponding hue/chroma/luminance coordinates and displays them in HCL space with one dimension collapsed. The collapsed dimension is the luminance for qualitative palettes and the hue for sequential/diverging palettes.
- `specplot()` also converts the colors to hue/chroma/luminance coordinates but draws the resulting spectrum in a line plot.

For the qualitative `"Dark 3"` palette from above the following plots can be obtained (see Figure 5).

```
R> demoplot(q4, "bar")
R> hclplot(q4)
R> specplot(q4, type = "o")
```

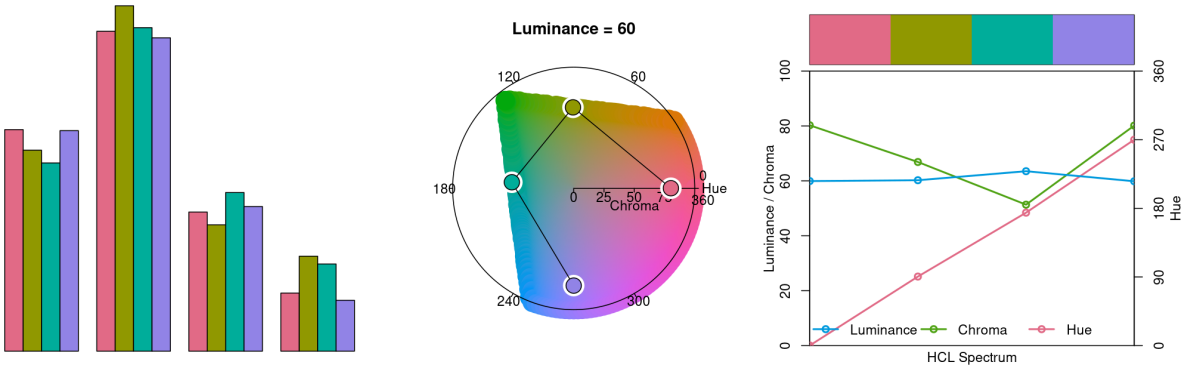

Figure 5: Palette visualization and assessment for the `qualitative_hcl(4, "Dark 3")` palette. Left: Demo bar plot. Center: Hue-chroma plane at fixed $L = 60$ in HCL space. Right: HCL spectrum with linearly changing hue (around color wheel), almost constant chroma, and constant luminance.

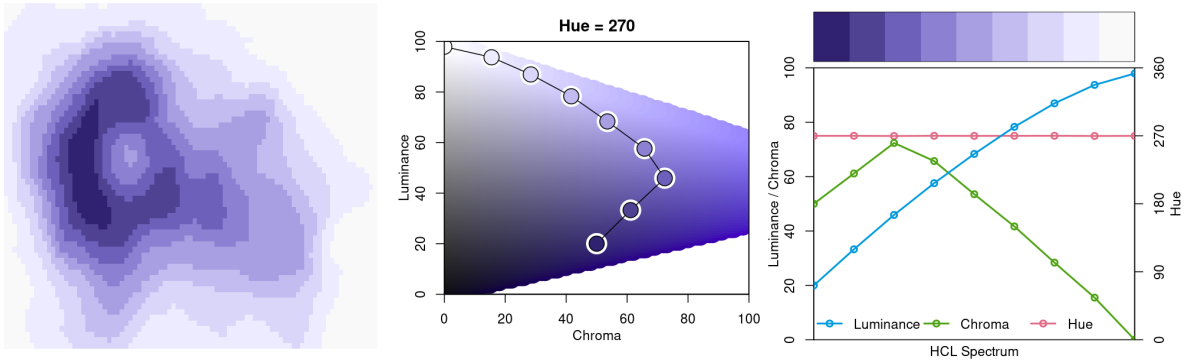

Figure 6: Palette visualization and assessment for the `sequential_hcl(9, "Purples 3")` palette. Left: Demo heatmap. Center: Chroma-luminance plane at fixed $H = 270$ in HCL space. Right: HCL spectrum with constant hue, triangular chroma, and increasing luminance.

The bar plot is used as a typical application for a qualitative palette (in addition to the time series and density plots used above). The other two displays show that luminance is (almost) constant in the palette while the hue changes linearly along the color "wheel" (from degree 0 to 270). Ideally, chroma would have also been constant to completely balance the colors. However, at this luminance the maximum chroma differs across hues so that the palette is fixed up to use less chroma for the yellow and green elements.

Note also that in a bar plot areas are shaded (and not just points or lines) so that lighter colors would be preferable. In the density plot in Figure 4 this was achieved through semi-transparency. Alternatively, luminance could be increased as is done in the `"Pastel 1"` or `"Set 3"` palettes.

Subsequently, the same types of assessment are carried out in Figure 6 for the sequential `"Purples 3"` palette as employed above.

```
R> s9 <- sequential_hcl(9, "Purples 3")
R> demoplot(s9, "heatmap")
R> hclplot(s9)
```

```
R> specplot(s9, type = "o")
```

In Figure 6, a heatmap (based on the well-known Maunga Whau volcano data) is used as a typical application for a sequential palette. The elevation of the volcano is brought out clearly, using dark colors to give emphasis to higher elevations. The other two displays show that hue is constant in the palette while luminance and chroma vary. Luminance increases monotonically from dark to light (as required for a proper sequential palette). Chroma is triangular-shaped which allows the viewer to better distinguish the middle colors in the palette when compared to a monotonic chroma trajectory.

3. Color spaces: S4 classes and utilities

At the core of the **colorspace** package are various utilities for computing with color spaces (Wikipedia 2020d), as the name of the package conveys. Thus, the package helps to map various three-dimensional representations of color to each other (Ihaka 2003). A particularly important mapping is the one from the perceptually-based and device-independent color model HCL (hue-chroma-luminance) to standard red-green-blue (sRGB) which is the basis for color specifications in many systems based on the corresponding hexadecimal (or simply hex) codes (Wikipedia 2020i), e.g., in HTML but also in R. For completeness further standard color models are included as well in the package. Their connections are illustrated in Figure 7. Color models that are (or try to be) perceptually-based are displayed with circles and models that are not are displayed with rectangles.

3.1. Implemented color spaces

The color spaces, implemented in **colorspace**, along with their corresponding S4 classes and eponymous class constructors, are:

- `RGB()` for the classic red-green-blue color model, which mixes three primary colors with different intensities to obtain a spectrum of colors. The advantage of this color model is (or was) that it corresponded to how computer and TV screens generated colors, hence it was widely adopted and still is the basis for color specifications in many systems. For example, hex color codes are employed in HTML but also in R. However, the RGB model also has some important drawbacks: It does not take into account the output device properties, it is not perceptually uniform (a unit step within RGB does not produce a constant perceptual change in color), and it is unintuitive for humans to specify colors (say brown or pink) in this space. See Wikipedia (2020g) for more details.
- `sRGB()` addresses the issue of device dependency by adopting a so-called gamma correction. Therefore, the gamma-corrected standard RGB (sRGB), as opposed to the linearized RGB above, is a good model for specifying colors in software and for hardware. But it is still unintuitive for humans to work directly with this color space. Therefore, sRGB is a good place to end up in a color space manipulation but it is not a good place to start. See Wikipedia (2020h) for more details.
- `HSV()` is a simple transformation of either the sRGB or the RGB space that tries to capture the perceptual axes: *hue* (dominant wavelength, the type of color), *saturation* (colorfulness), and *value* (brightness, i.e., light vs. dark). Unfortunately, the three axes

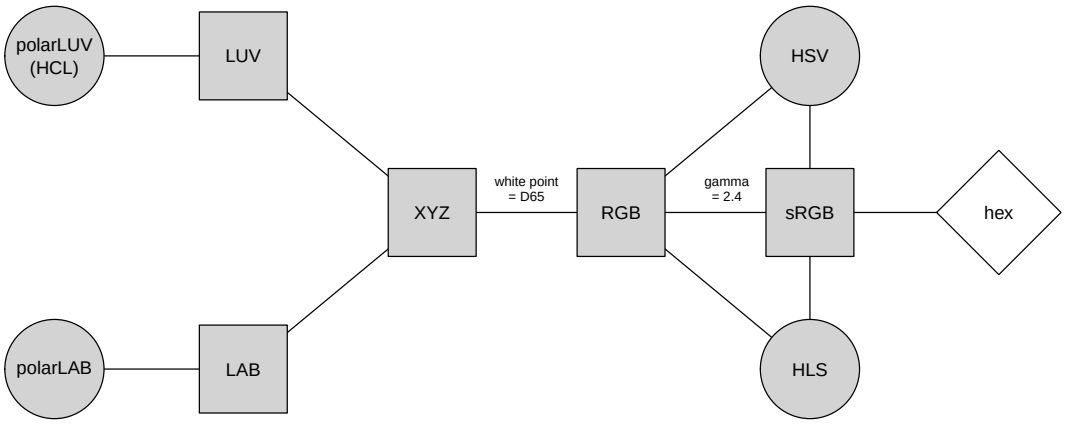

Figure 7: Relationships among three-dimensional color spaces implemented in **colorspace**. Color models that are (or try to be) perceptually-based are displayed with circles, other color models with rectangles.

in the HSV model are confounded so that, e.g., brightness changes dramatically with hue. See Wikipedia (2020f) for more details.

- `HLS()` (hue-lightness-saturation) is another transformation of either sRGB or RGB that tries to capture the perceptual axes. It does a somewhat better job but the dimensions are still strongly confounded. See Wikipedia (2020f) for more details.
- `XYZ()` was established by the CIE (Commission Internationale de l'Eclairage) based on psychophysical experiments with human subjects. It provides a unique triplet of XYZ values, coding the standard observer's perception of the color. It is device-independent but it is not perceptually uniform and the XYZ coordinates have no intuitive meaning. See Wikipedia (2020a) for more details.
- `LUV()` and `LAB()` were therefore proposed by the CIE as perceptually uniform color spaces where the former is typically preferred for emissive technologies (such as screens and monitors) whereas the latter is usually preferred when working with dyes and pigments. The L coordinate in both spaces has the same meaning and captures luminace (light-dark contrasts). Both the U and V coordinates as well as the A and B coordinates measure positions on red/green and yellow/blue axes, respectively, albeit in somewhat different ways. While this corresponds to how human color vision likely evolved (see the next section), these two color models still not correspond to perceptual axes that humans use to describe colors. See Wikipedia (2020c,b) for more details.
- `polarLUV()` and `polarLAB()` take polar coordinates in the UV plane and AB plane, respectively. Specifically, the polar coordinates of the LUV model are known as the HCL (hue-chroma-luminance) model (see Wikipedia 2020e, which points out that the LAB-based polar coordinates are also sometimes referred to as HCL). The HCL model captures the human perceptual axes very well without confounding effects as in the HSV or HLS approaches. (More details follow below.)

All S4 classes for color spaces inherit from a virtual class '`color`' which is internally always represented by matrices with three columns (corresponding to the three dimensions).

Note that since the inception of the color space conversion tools within **colorspace** (in C, Ihaka 2003) other R tools for this purpose became available, notably `grDevices::convertColor()`

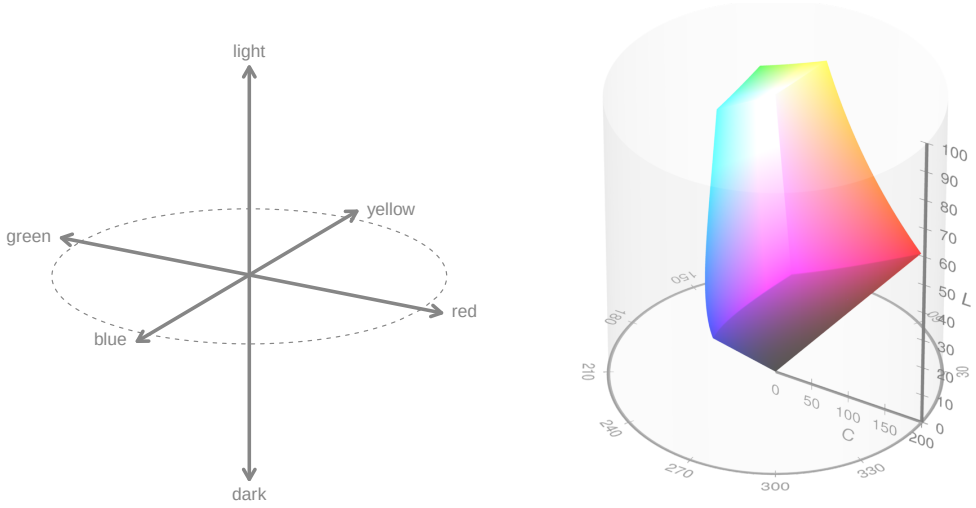

Figure 8: Visualization of axes capturing human color vision (left) and the corresponding HCL color model (right).

(in high-level R, R Core Team 2020) and `farver::convert_colour()` (in C++, Pedersen, Nicolae, and François 2020). For many basic color conversion purposes the **colorspace** package and these alternatives are essentially equally suitable (see the discussion in Zeileis, Gaslam, Murrell, and Pedersen 2018). For more complex conversions, including different chromatic adaptation algorithms, a more comprehensive color science approach is implemented in the R package **colorscience** (Gama and Davis 2018). Finally, base R also provides `grDevices::hcl()` for mapping HCL representations to hex codes.

To make the **colorspace** package self-contained and exactly backward compatible, the C code in **colorspace** is still used as the basis for all color space conversions.

3.2. Human color vision and the HCL color model

It has been hypothesized that human color vision has evolved in three distinct stages:

1. Perception of light/dark contrasts (monochrome only).
2. Yellow/blue contrasts (usually associated with our notion of warm/cold colors).
3. Green/red contrasts (helpful for assessing the ripeness of fruit).

See Kaiser and Boynton (1996), Knoblauch (2002), Ihaka (2003), Lumley (2006), Zeileis *et al.* (2009) for more details and references. Thus, colors can be described using a 3-dimensional space as shown in the left panel of Figure 8. However, for describing colors in such a space, it is more natural for humans to employ polar coordinates in the color plane (yellow/blue vs. green/red, visualized by the dashed circle in Figure 8) plus a third light/dark axis. Hence, color models that attempt to capture these perceptual axes are also called perceptually-based color spaces. As already argued above, the HCL model captures these dimensions very well, calling them: *hue*, *chroma*, and *luminance*. The corresponding sRGB gamut, i.e., the HCL colors that can also be represented in sRGB, is visualized in the right panel of Figure 8 (by Horvath and Lipka 2016). An animated version of the same plot is provided online by Horvath and Lipka (2017).

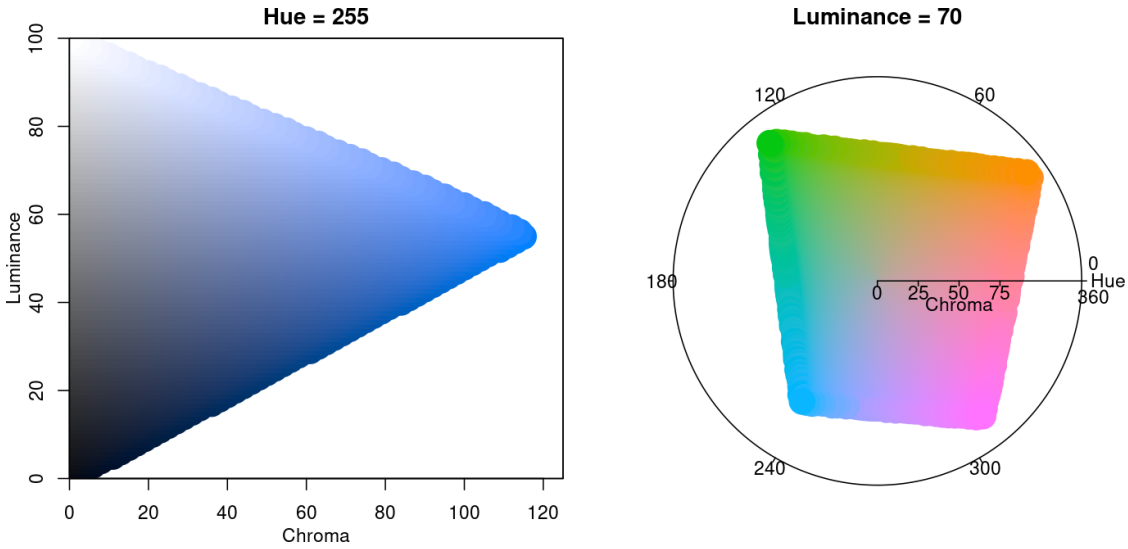

Figure 9: Vertical (left) and horizontal (right) slices of the HCL space yielding a chroma-luminance plane for given hue and a hue-chroma plane for given luminance, respectively.

The shape of the HCL space is a distorted double cone which is seen best by looking at vertical slices, i.e., chroma-luminance planes for given hues. For example, the left panel in Figure 9 depicts the chroma-luminance plane for a certain blue (hue = 255). Along with luminance the colors change from dark to light. With increasing chroma the colors become more colorful, where the highest chroma is possible for intermediate luminance.

As some colors are relatively dark (e.g., blue and red assume their maximum chroma for relatively low luminances) while others are relatively light (e.g., yellow and green), horizontal slices of hue-chroma planes for given hue have somewhat irregular shapes. The right panel in Figure 9 shows such a hue-chroma plane for moderately light colors (luminance = 70). At that luminance, green and orange can become much more colorful compared to blue or red.

3.3. Utilities

Several utilities are available for working with the S4 classes implementing the color spaces listed above.

- `as()` method: Convert a '`color`' object to the various color spaces, e.g., `as(x, "sRGB")`.
- `coords()`: Extract the three-dimensional coordinates pertaining to the current '`color`' class.
- `hex()`: Convert a '`color`' object to '`sRGB`' and code in a hex string that can be used within R plotting functions.
- `hex2RGB()`: Convert a given hex color string to an '`sRGB`' color object which can also be coerced to other color spaces.
- `readRGB()` and `readhex()` can read text files into '`color`' objects, either from RGB coordinates or hex color strings.
- `writehex()`: Write hex color strings to a text file.
- `whitepoint()`: Query and change the so-called white point employed in conversions from CIE XYZ to RGB. Defaults to D65 that has been specified by the CIE to approximate daylight (Poynton 2009, FAQ 15).

3.4. Illustration of basic colorspace functionality

As an example a vector of colors `x` can be specified in the HCL (or polar LUV) model:

```
R> (x <- polarLUV(L = 70, C = 50, H = c(0, 120, 240)))

      L  C   H
[1,] 70 50   0
[2,] 70 50 120
[3,] 70 50 240
```

The resulting three colors are pastel red (hue = 0), green (hue = 120), and blue (hue = 240) with moderate chroma and luminance. For display in other systems an sRGB representation might be needed:

```
R> (y <- as(x, "sRGB"))

             R         G         B
[1,] 0.8931564 0.5853740 0.6465459
[2,] 0.5266113 0.7224335 0.4590469
[3,] 0.4907804 0.6911937 0.8673877
```

The displayed coordinates can also be extracted as numeric matrices by `coords(x)` or `coords(y)`. We can also, for example, coerce from sRGB to HSV:

```
R> as(y, "HSV")

            H         S         V
[1,] 348.0750 0.3446008 0.8931564
[2,] 104.6087 0.3645825 0.7224335
[3,] 208.0707 0.4341857 0.8673877
```

For display in many systems (including R itself) hex color codes based on the sRGB coordinates can be created:

```
R> hex(x)

[1] "#E495A5" "#86B875" "#7DB0DD"
```

4. HCL-based color palettes

As motivated in the previous section, the HCL space is particularly useful for specifying individual colors and color palettes, as its three axes match those of the human visual system very well. Therefore, the **colorspace** package provides three palette functions based on the HCL model: `qualitative_hcl()`, `sequential_hcl()`, and `diverging_hcl()`. Their construction principles are exemplified in Figure 10 and explained in more detail below. The desaturated

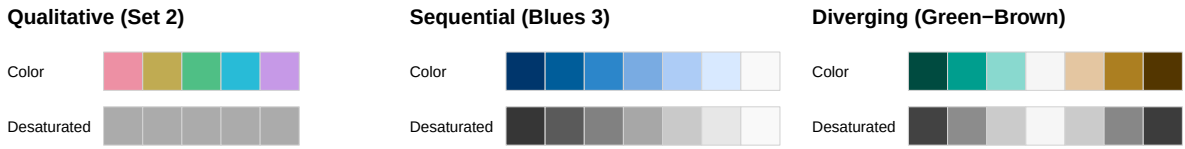

Figure 10: Examples of palette types in **colorspace**. Qualitative palettes are balanced towards the same luminance level while sequential and diverging palettes go from dark to light and/or vice versa, respectively.

palettes in the second row of Figure 10 bring out clearly that luminance differences (light-dark contrasts) are crucial for sequential and diverging palettes while qualitative palettes are balanced at the same luminance.

To facilitate obtaining good sets of colors, HCL parameter combinations that yield useful palettes are accessible by name. These can be listed using the function `hcl_palettes()`:

```
R> hcl_palettes()

HCL palettes

Type:  Qualitative
Names: Pastel 1, Dark 2, Dark 3, Set 2, Set 3, Warm, Cold, Harmonic,
       Dynamic

Type:  Sequential (single-hue)
Names: Grays, Light Grays, Blues 2, Blues 3, Purples 2, Purples 3, Reds 2,
       Reds 3, Greens 2, Greens 3, Oslo

Type:  Sequential (multi-hue)
Names: Purple-Blue, Red-Purple, Red-Blue, Purple-Orange, Purple-Yellow,
       Blue-Yellow, Green-Yellow, Red-Yellow, Heat, Heat 2,
       Terrain, Terrain 2, Viridis, Plasma, Inferno, Dark Mint,
       Mint, BluGrn, Teal, TealGrn, Emrld, BluYl, ag_GrnYl, Peach,
       PinkYl, Burg, BurgYl, RedOr, OrYel, Purp, PurpOr, Sunset,
       Magenta, SunsetDark, ag_Sunset, BrwnYl, YlOrRd, YlOrBr,
       OrRd, Oranges, YlGn, YlGnBu, Reds, RdPu, PuRd, Purples,
       PuBuGn, PuBu, Greens, BuGn, GnBu, BuPu, Blues, Lajolla,
       Turku, Hawaii, Batlow

Type:  Diverging
Names: Blue-Red, Blue-Red 2, Blue-Red 3, Red-Green, Purple-Green,
       Purple-Brown, Green-Brown, Blue-Yellow 2, Blue-Yellow 3,
       Green-Orange, Cyan-Magenta, Tropic, Broc, Cork, Vik, Berlin,
       Lisbon, Tofino
```

To inspect the HCL parameter combinations for a specific palette simply include the `palette` name where upper- vs. lower-case, spaces, etc. are ignored for matching the label, e.g., `"set2"` matches `"Set 2"`:

```
R> hcl_palettes(palette = "set2")

HCL palette
Name: Set 2
Type: Qualitative
Parameter ranges:
 h1 h2 c1 c2 l1 l2 p1 p2 cmax fixup
  0 NA 60 NA 70 NA NA NA   NA  TRUE
```

To compute the actual color hex codes (representing sRGB coordinates) based on these HCL parameters, the functions `qualitative_hcl()`, `sequential_hcl()`, and `diverging_hcl()` can be used which are described in more detail in the following sections. Either all parameters can be specified "by hand" through the HCL parameters, an entire palette can be specified "by name", or the name-based specification can be modified by a few HCL parameters. In case of the HCL parameters, either a vector-based specification such as `h = c(0, 270)` or individual parameters `h1 = 0` and `h2 = 270` can be used.

The first three of the following commands lead to equivalent output. The fourth command yields a modified set of colors (lighter due to a luminance of 80 instead of 70).

```
R> qualitative_hcl(4, h = c(0, 270), c = 60, l = 70)

[1] "#ED90A4" "#ABB150" "#00C1B2" "#ACA2EC"

R> qualitative_hcl(4, h1 = 0, h2 = 270, c1 = 60, l1 = 70)

[1] "#ED90A4" "#ABB150" "#00C1B2" "#ACA2EC"

R> qualitative_hcl(4, palette = "set2")

[1] "#ED90A4" "#ABB150" "#00C1B2" "#ACA2EC"

R> qualitative_hcl(4, palette = "set2", l = 80)

[1] "#FFACBF" "#C6CD70" "#32DDCD" "#C7BEFF"
```

4.1. Qualitative palettes

As suggested by Ihaka (2003), `qualitative_hcl()` distinguishes the underlying categories by a sequence of hues while keeping both chroma and luminance constant, to give each color in the resulting palette the same perceptual weight. Thus, `h` should be a pair of hues (or equivalently `h1` and `h2` can be used) with the starting and ending hue of the palette. Then, an equidistant sequence between these hues is employed, by default spanning the full color wheel (i.e., the full 360 degrees). Chroma `c` (or equivalently `c1`) and luminance `l` (or equivalently `l1`) are constants. Finally, `fixup` indicates whether colors with out-of-range coordinates should be corrected (as illustrated in Figure 5).

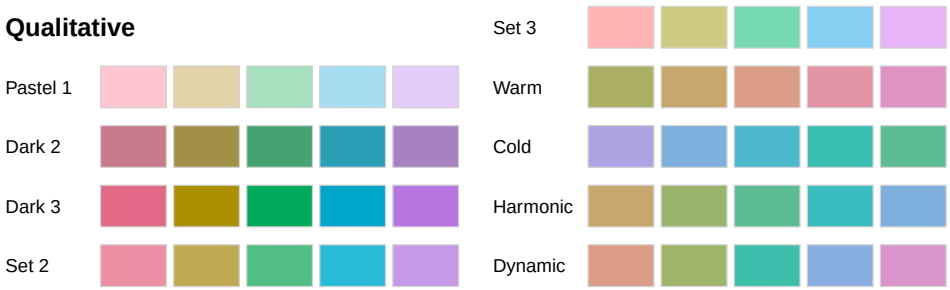

Figure 11: Prespecified qualitative HCL palettes available in `qualitative_hcl()` in **colorspace**.

Figure 11 shows the named palettes available in the `qualitative_hcl()` function. The first five palettes are close to the **ColorBrewer.org** palettes of the same name (Harrower and Brewer 2003). They employ different levels of chroma and luminance and, by default, span the full hue range. The remaining four palettes are taken from Ihaka (2003). They are based on the same chroma (50) and luminance (70) but the hue is restricted to different intervals.

```
R> hcl_palettes("qualitative", plot = TRUE, nrow = 5)
```

When palettes are employed for shading areas in statistical displays (e.g., in bar plots, pie charts, or regions in maps), lighter colors (with moderate chroma and high luminance) such as `"Pastel 1"` or `"Set 3"` are typically less distracting. By contrast, when coloring points or lines, more flashy colors (with high chroma) are often required: On a white background a moderate luminance as in `"Dark 2"` or `"Dark 3"` usually works better while on a black/dark background the luminance should be higher as in `"Set 2"`. Some examples with demo graphics are provided in Section 5.

4.2. Sequential palettes (single-hue)

As suggested by Zeileis *et al.* (2009), `sequential_hcl()` codes the underlying numeric values by a monotonic sequence of increasing (or decreasing) luminance. Thus, the function's `l` argument should provide a vector of length 2 with starting and ending luminance (equivalently, `l1` and `l2` can be used). Without chroma (i.e., `c = 0`), this simply corresponds to a grayscale palette like `gray.colors()`, see `"Grays"` and `"Light Grays"` in Figure 12.

For adding chroma, a simple strategy would be to pick a single hue value (via `h` or `h1`) and then decrease chroma from some value (`c` or `c1`) to zero (i.e., gray) along with increasing luminance. This is already very effective for bringing out the extremes (a dark high-chroma color vs. a light gray), see `"Blues 2"`, `"Purples 2"`, `"Reds 2"`, and `"Greens 2"`.

For distinguishing colors in the center of the palette, two strategies can be employed: (a) Hue can be varied as well by specifying an interval of hues in `h` (or beginning hue `h1` and ending hue `h2`). More details are provided in the next section. (b) Instead of a decreasing chroma, a triangular chroma trajectory can be employed from `c1` over `cmax` to `c2` (equivalently specified as a vector `c` of length 3). This yields high-chroma colors in the middle of the palette that are more easily distinguished from the dark and light extremes. See `"Blues 3"`, `"Purples 3"`, `"Reds 3"`, and `"Greens 3"` in Figure 12.

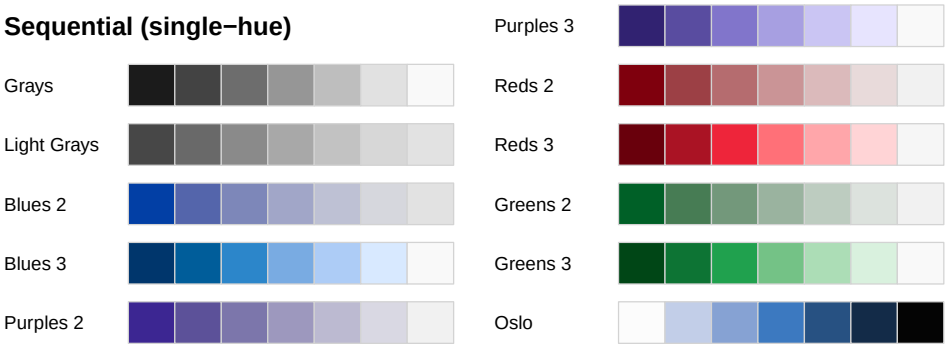

Figure 12: Prespecified sequential single-hue HCL palettes available in `sequential_hcl()` in **colorspace**.

Instead of employing linear trajectories in the chroma or luminance coordinates, some palettes employ a power transformation of the chroma and/or luminance trajectory. Either a vector `power` of length 2 or separate `p1` (for chroma) and `p2` (for luminance) can be specified. If the latter is missing, it defaults to the former.

```
R> hcl_palettes("sequential (single-hue)", n = 7, plot = TRUE, nrow = 6)
```

All except the last are inspired by the **ColorBrewer.org** palettes with the same base name (Harrower and Brewer 2003) but restricted to a single hue only. They are intended for a white/light background. The last palette (`"Oslo"`) is taken from the scientific color maps of Crameri (2018) and is intended for a black/dark background and hence the order is reversed starting from a light blue (not a light gray).

To distinguish many colors in a sequential palette it is important to have a strong contrast on the luminance axis, possibly enhanced by an accompanying pronounced variation in chroma. When only a few colors are needed (e.g., for coding an ordinal categorical variable with few levels) then a lower luminance contrast may suffice.

4.3. Sequential palettes (multi-hue)

To not only bring out extreme colors in a sequential palette but also better distinguish middle colors it is a common strategy to employ a sequence of hues. Thus, the basis of such a palette is still a monotonic luminance sequence as above (combined with a monotonic or triangular chroma sequence). But rather than using a single hue, an interval of hues in `h` (or beginning hue `h1` and ending hue `h2`) can be specified.

`sequential_hcl()` allows combined variations in hue (`h` and `h1`/`h2`, respectively), chroma (`c` and `c1`/`c2`/`cmax`, respectively), luminance (`l` and `l1`/`l2`, respectively), and power transformations for the chroma and luminance trajectories (`power` and `p1`/`p2`, respectively). This yields a broad variety of sequential palettes, including many that closely match other well-known color palettes. Figure 13 shows all the named multi-hue sequential palettes in **colorspace**:

```
R> hcl_palettes("sequential (multi-hue)", n = 7, plot = TRUE)
```

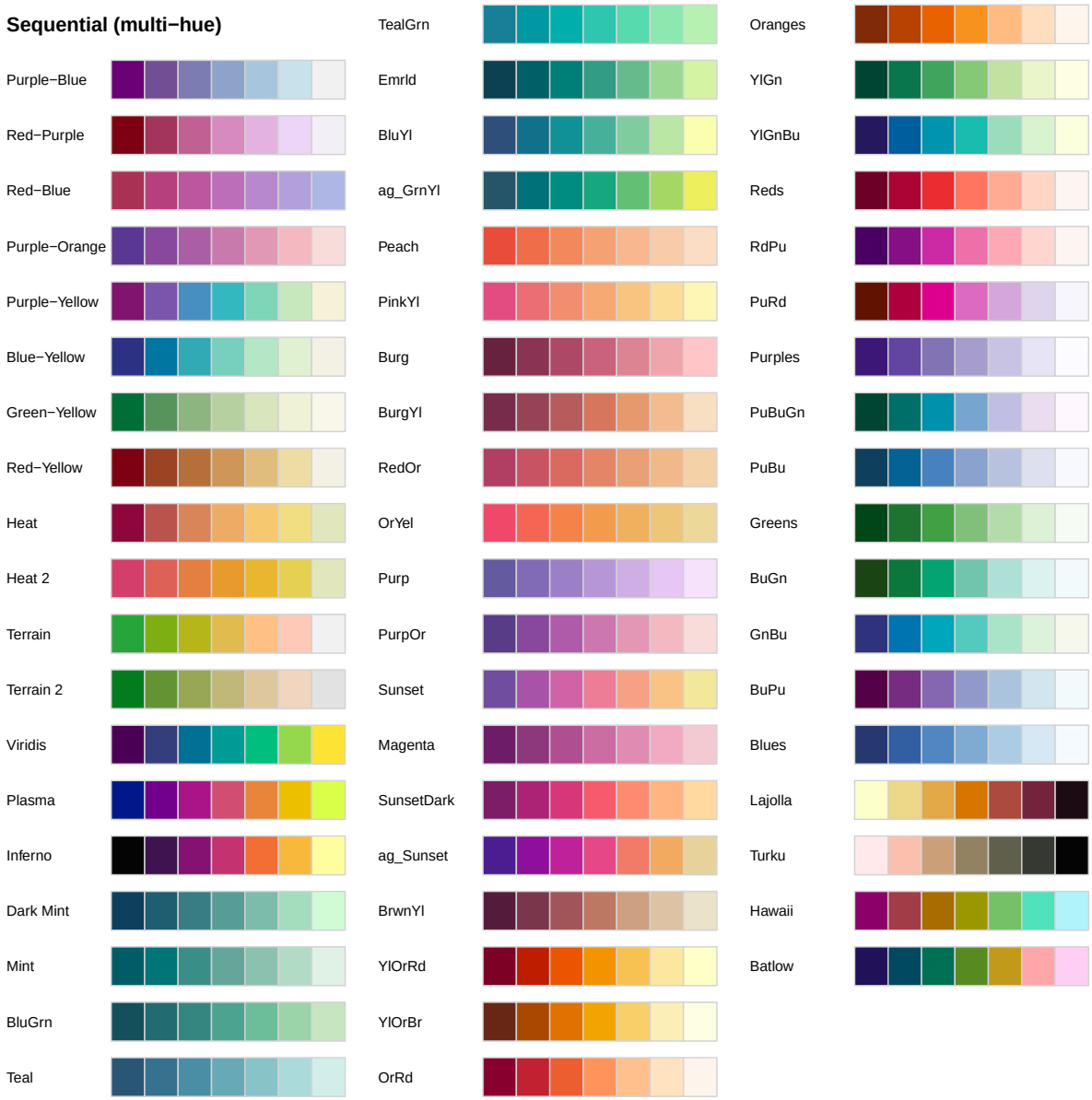

Figure 13: Prespecified sequential multi-hue HCL palettes available in `sequential_hcl()` in **colorspace**.

- `"Purple-Blue"` to `"Terrain 2"` are various palettes created during the development of **colorspace**, e.g., by Zeileis *et al.* (2009) or Stauffer *et al.* (2015) among others.
- `"Viridis"` to `"Inferno"` closely match the palettes that Smith and Van der Walt (2015) developed for **matplotlib** and that gained popularity recently.
- `"Dark Mint"` to `"BrwnYl"` closely match palettes provided in **CARTO** (CARTO 2019).
- `"YlOrRd"` to `"Blues"` closely match **ColorBrewer.org** palettes (Harrower and Brewer 2003).
- `"Lajolla"` to `"Batlow"` closely match the scientific color maps of the same name by Crameri (2018) and the first two of these are intended for a black/dark background.

Figure 14: Prespecified diverging HCL palettes available in `diverging_hcl()` in **colorspace**.

Note that the palettes differ substantially in the amount of chroma and luminance contrasts. For example, many palettes go from a dark high-chroma color to a neutral low-chroma color (e.g., `"Reds"`, `"Purples"`, `"Greens"`, `"Blues"`) or even light gray (e.g., `"Purple-Blue"`). But some palettes also employ relatively high chroma throughout the palette (e.g., the **viridis** and many **CARTO** palettes). To emphasize the extremes the former strategy is typically more suitable while the latter works better if all values along the sequence should receive some more perceptual weight.

4.4. Diverging palettes

`diverging_hcl()` codes the underlying numeric values by a triangular luminance sequence with different hues in the left and in the right "arms" of the palette. Thus, it can be seen as a combination of two sequential palettes with some restrictions: (a) a single hue is used for each arm of the palette, (b) chroma and luminance trajectory are balanced between the two arms, (c) the neutral central value has zero chroma. To specify such a palette a vector of two hues `h` (or equivalently `h1` and `h2`), either a single chroma value `c` (or `c1`) or a vector of two chroma values `c` (or `c1` and `cmax`), a vector of two luminances `l` (or `l1` and `l2`), and power parameter(s) `power` (or `p1` and `p2`) are used. For more flexible diverging palettes without the restrictions above (and consequently more parameters) see the `divergingx_hcl()` palettes introduced below.

Figure 14 shows all such diverging palettes that have been named in **colorspace**:

```
R> hcl_palettes("diverging", n = 7, plot = TRUE, nrow = 10)
```

- `"Blue-Red"` to `"Cyan-Magenta"` have been developed for **colorspace** starting from Zeileis *et al.* (2009), taking inspiration from various other palettes, including more balanced and simplified versions of several **ColorBrewer.org** palettes (Harrower and Brewer 2003).
- `"Tropic"` closely matches the palette of the same name from **CARTO** (CARTO 2019).

- `"Broc"` to `"Vik"` and `"Berlin"` to `"Tofino"` closely match the scientific color maps of the same name by Crameri (2018), where the first three are intended for a white/light background and the other three for a black/dark background.

When choosing a particular palette for a display similar considerations apply as for the sequential palettes. Thus, large luminance differences are important when many colors are used while smaller luminance contrasts may suffice for palettes with fewer colors etc.

4.5. Construction details

Table 1 summarizes which types of trajectories (*constant*, *linear*, *triangular*) are used for the three HCL coordinates (hue H, chroma C, luminance L) to construct the different types of palettes (*qualitative*, *sequential*, and *diverging*).

As emphasized in Figure 10, luminance is probably the most important property for defining the type of palette. It is constant for qualitative palettes, monotonic for sequential palettes (linear or a power transformation), or uses two monotonic trajectories (linear or a power transformation) diverging from the same neutral value.

Hue trajectories are also rather intuitive and straightforward for the three different types of palettes (constant vs. linear). However, chroma trajectories are probably the most complicated and least obvious from the examples above. Hence, the exact mathematical equations underlying the chroma trajectories are given in the following (i.e., using the parameters `c1`, `c2`, `cmax`, and `p1`, respectively) and are depicted in Figure 15. Analogous equations apply for the other two coordinates.

The trajectories are functions of the *intensity* $i \in [0, 1]$ where 1 corresponds to the full intensity:

$$\text{Constant: } c_1 \tag{1}$$

$$\text{Linear: } c_2 - (c_2 - c_1) \cdot i \tag{2}$$

$$\text{Triangular: } \begin{cases} c_2 - (c_2 - c_{\max}) \cdot \frac{i}{j} & \text{if } i \leq j \\ c_{\max} - (c_{\max} - c_1) \cdot \frac{i-j}{1-j} & \text{if } i > j \end{cases} \tag{3}$$

where j is the intensity at which $c_{\max}$ is assumed. It is constructed such that the slope to the left is the negative of the slope to the right of j:

$$j = \left(1 + \frac{|c_{\max} - c_1|}{|c_{\max} - c_2|}\right)^{-1}$$

Instead of using a linear intensity i going from 1 to 0, one can replace i with i^{p_1} in Equations 1–3. This then leads to power-transformed curves that add or remove chroma more slowly or more quickly depending on whether the power parameter p_1 is < 1 or > 1.

The three types of trajectories are also depicted in Figure 15. Note that full intensity $i = 1$ is on the left and zero intensity $i = 0$ is on the right of each panel. The concrete parameters are:

- Constant: `c1 = 80`.
- Linear: `c1 = 80`, `c2 = 10`, `p1 = 1` (black) vs. `p1 = 1.6` (gray).
- Triangular: `c1 = 60`, `cmax = 80`, `c2 = 10`, `p1 = 1` (black) vs. `p1 = 1.6` (gray).

Type	H	C	L
Qualitative	Linear	Constant	Constant
Sequential	Constant (single-hue) *or* Linear (multi-hue)	Linear (+ power) *or* Triangular (+ power)	Linear (+ power)
Diverging	Constant (2×)	Linear (+ power) *or* Triangular (+ power)	Linear (+ power)

Table 1: Types of trajectories used for the HCL coordinates to construct qualitative, sequential, and diverging palettes, see Equations 1–3.

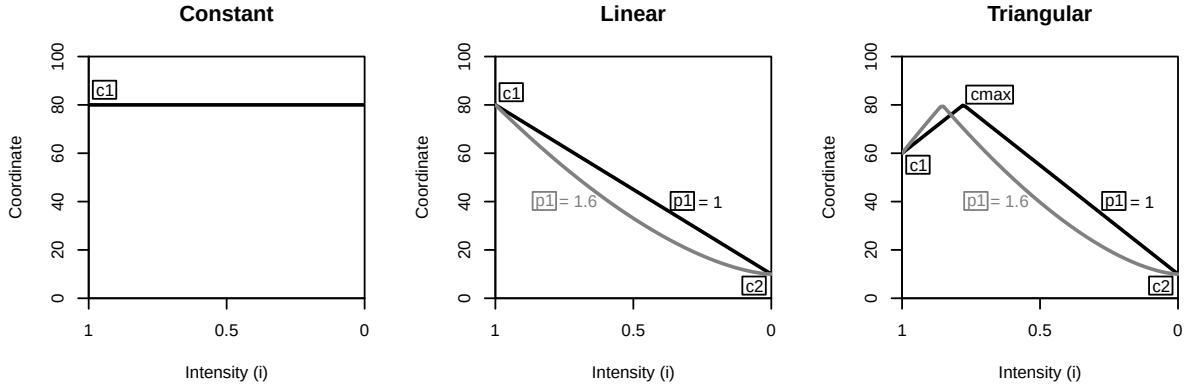

Figure 15: Types of trajectories to construct HCL color palettes, exemplified for the chroma coordinates, see Equations 1–3.

Further discussion of these trajectories and how they can be visualized and assessed for a given color palette is provided in Section 5.

4.6. Registering your own palettes

The `hcl_palettes()` already come with a wide range of predefined palettes to which customizations can be easily added. However, it might also be convenient to register a custom palette so that it can subsequently be reused with a new dedicated name. This is supported by adding a `register` argument once to a call to `qualitative_hcl()`, `sequential_hcl()`, or `diverging_hcl()`:

```
R> qualitative_hcl(3, palette = "set2", l = 80, register = "myset")
```

The new palette is then included in `hcl_palettes()`:

```
R> hcl_palettes("Qualitative")

HCL palettes

Type:  Qualitative
Names: Pastel 1, Dark 2, Dark 3, Set 2, Set 3, Warm, Cold, Harmonic,
       Dynamic, myset
```

The palette can be used subsequently in `qualitative_hcl()` as well as the qualitative **ggplot2** color scales (see Section 2.3), e.g.,

```
R> qualitative_hcl(4, palette = "myset")

[1] "#FFACBF" "#C6CD70" "#32DDCD" "#C7BEFF"
```

Remarks:

- The number of colors in the palette that was used during registration is not actually stored and can be modified subsequently. The same holds for arguments `alpha` and `rev`.
- When registering a new palette with a previously-used name, the old palette gets overwritten. We recommend to not overwrite the palettes that are predefined in the package (albeit technically possible).
- The registration of a palette is only stored for the current session. When R is restarted and/or the **colorspace** package reloaded, only the predefined palettes from the package are available. Thus, to make a palette permanently available a registration R code like `colorspace::qualitative_hcl(3, palette = "set2", l = 80, register = "myset")` can be placed in your `.Rprofile` or similar startup scripts.

4.7. Flexible diverging palettes

The `divergingx_hcl()` function provides more flexible diverging palettes by simply calling `sequential_hcl()` twice with prespecified sets of hue, chroma, and luminance parameters. Thus, it does not pose any restrictions that the two "arms" of the palette need to be balanced and also may go through a non-gray neutral color (typically light yellow). Consequently, the chroma/luminance paths can be rather unbalanced.

Figure 16 shows all such flexible diverging palettes that have been named in **colorspace**:

```
R> divergingx_palettes(n = 7, plot = TRUE, nrow = 10)
```

- `"ArmyRose"` to `"Tropic"` closely match the palettes of the same name from **CARTO** (CARTO 2019).
- `"PuOr"` to `"Spectral"` closely match the palettes of the same name from **ColorBrewer.org** (Harrower and Brewer 2003).
- `"Zissou 1"` closely matches the palette of the same name from **wesanderson** (Ram and Wickham 2018).
- `"Cividis"` closely matches the palette of the same name from the **viridis** family (Garnier 2018). Note that despite having two "arms" with blue vs. yellow colors and a low-chroma center color, this is probably better classified as a sequential palette due to the monotonic chroma going from dark to light. (See Section 4.8 for more details.)
- `"Roma"` closely matches the palette of the same name by Crameri (2018).

Typically, the more restricted `diverging_hcl()` palettes should be preferred because they are more balanced. However, by being able to go through light yellow as the neutral color warmer diverging palettes are available.

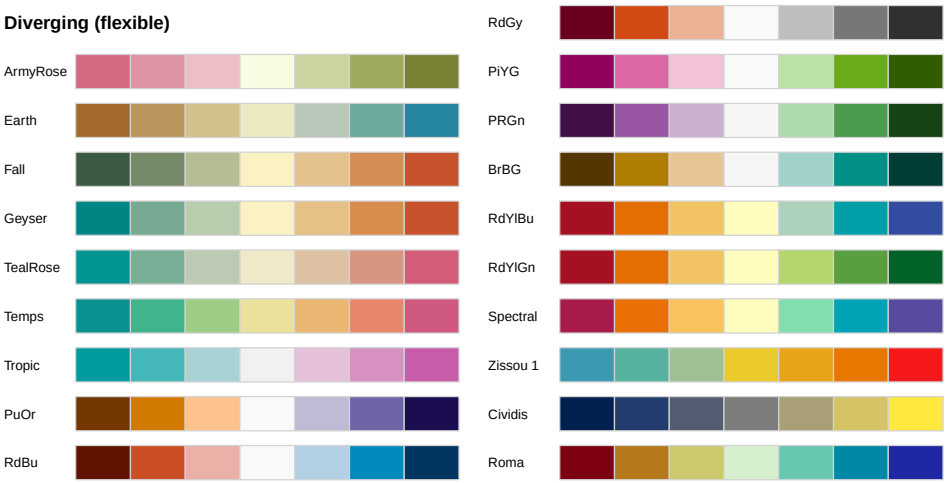

Figure 16: Prespecified flexible diverging HCL palettes available in `divergingx_hcl()` in **colorspace**.

4.8. Approximating palettes from other packages

The flexible specification of HCL-based color palettes in **colorspace** allows one to closely approximate color palettes from various other packages:

- **ColorBrewer.org** (Harrower and Brewer 2003) as provided by the R package **RColorBrewer** (Neuwirth 2014). See `demo("brewer", package = "colorspace")`.
- **CARTO** colors (CARTO 2019) as provided by the R package **rcartocolor** (Nowosad 2019). See `demo("carto", package = "colorspace")`.
- The viridis palettes of Smith and Van der Walt (2015) developed for **matplotlib**, as provided by the R package **viridis** (Garnier 2018). See `demo("viridis", package = "colorspace")`.
- The scientific color maps of Crameri (2018) as provided by the R package **scico** (Pedersen and Crameri 2020). See `demo("scico", package = "colorspace")`.

The graphics resulting from the demos can also be viewed online at `http://colorspace.R-Forge.R-project.org/articles/approximations.html`.

Figure 17 shows a selection of such approximations using `specplot()` (see also Section 5.2) for two blue/green/yellow palettes (namely `RColorBrewer::brewer.pal(7, "YlGnBu")` and `viridis::viridis(7)`) and two purple/red/yellow palettes (namely `rcartocolor::carto_pal(7, "ag_Sunset")` and `viridis::plasma(7)`). Each panel compares the hue, chroma, and luminance trajectories of the original palettes (top swatches, solid lines) and their HCL-based approximations (bottom swatches, dashed lines). The palettes are not identical but very close for most colors. Note also that the chroma trajectories from the HCL palettes (green dashed lines) have some kinks which are due to fixing HCL coordinates at the boundaries of admissible RGB colors.

Furthermore, Figure 17 illustrates what sets the viridis palettes apart from other sequential palettes. While the hue and luminance trajectories of `"Viridis"` and `"YlGnBu"` are very similar, the chroma trajectories differ: While lighter colors (with high luminance) have low chroma for `"YlGnBu"`, they have increasing chroma for `"Viridis"`. Similarly, `"ag_Sunset"`

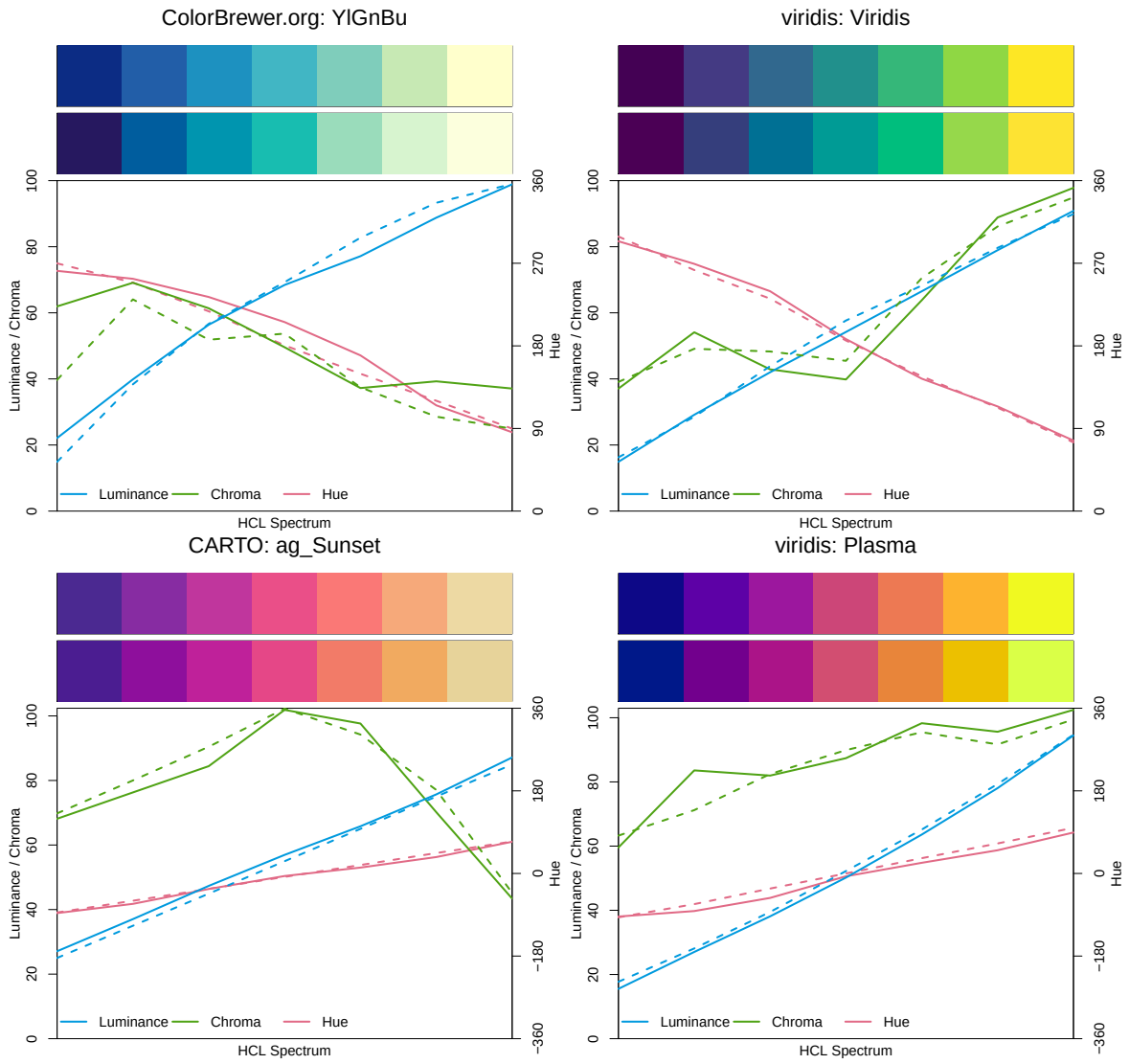

Figure 17: HCL spectrum of four palettes taken from **ColorBrewer.org**, **CARTO**, and **viridis** (top swatches, solid lines) along with their HCL-based approximations (bottom swatches, dashed lines).

and `"Plasma"` have similar hue and luminance trajectories but different chroma trajectories. The result is that the viridis palettes have rather high chroma throughout which does not work as well for sequential palettes on a white/light background as all shaded areas convey high "intensity". However, they work better on a dark/black background (see Figure 28 on page 33). Also, they might be a reasonable alternative for qualitative palettes when grayscale printing should also work.

Another somewhat nonstandard palette from the viridis family is the cividis palette based on blue and yellow hues and hence safe for red-green deficient viewers. Figure 18 shows the corresponding `specplot()` along with an HCL-based approximation. This palette is unusual: The hue and chroma trajectories would suggest a diverging palette, as there are two "arms"

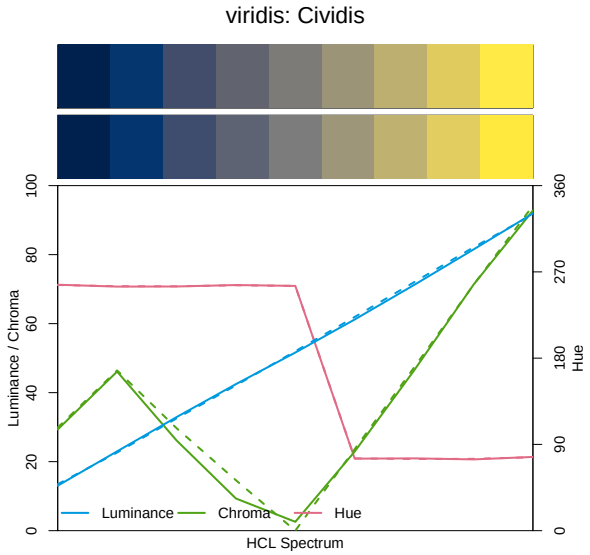

Figure 18: HCL spectrum of `viridis::cividis` (top swatch, solid lines) along with an HCL-based approximation (bottom swatch, dashed lines).

with different hues and a zero-chroma point in the center. However, the luminance trajectory clearly indicates a sequential palette as colors go monotonically from dark to light. Due to this unusual mixture the palette cannot be composed using the trajectories from Table 1.

However, the tools in **colorspace** can still be employed to easily reconstruct the palette. One strategy would be to set up the trajectories manually, using a linear luminance, piecewise linear chroma, and piecewise constant hue:

```
R> cividis_hcl <- function(n) {
+    i <- seq(1, 0, length.out = n)
+    hex(polarLUV(
+      L = 92 - (92 - 13) * i,
+      C = approx(c(1, 0.9, 0.5, 0), c(30, 50, 0, 95), xout = i)$y,
+      H = c(255, 75)[1 + (i < 0.5)]
+    ), fix = TRUE)
+  }
```

Instead of constructing the hex code from the HCL coordinates via `hex(polarLUV(L, C, H))` from **colorspace**, the base R function `hcl(H, C, L)` from **grDevices** could also be used.

In addition to manually setting up a dedicated function `cividis_hcl()`, it is possible to approximate the palette using `divergingx_hcl()` (see Section 4.7), e.g.,

```
R> divergingx_hcl(n,
+    h1 = 255, h2 = NA, h3 = 75,
+    c1 = 30, cmax1 = 47, c2 = 0, c3 = 95,
+    l1 = 13, l2 = 52, l3 = 92,
+    p1 = 1.1, p3 = 1.0
+  )
```

This uses a slight power transformation with `p1 = 1.1` in the blue arm of the palette but otherwise essentially corresponds to what `cividis_hcl()` does. For convenience the above parameters are already preregistered in `divergingx_hcl(n, palette = "Cividis")`.

4.9. HCL (and HSV) color palettes corresponding to base R palettes

To facilitate switching from base R palette functions to the HCL-based palettes above, **colorspace** provides a few convenience interfaces:

- `rainbow_hcl()`: Convenience interface to `qualitative_hcl()` for a HCL-based "rainbow" palette to replace the (in)famous `rainbow()` palette.
- `heat_hcl()`: Convenience interface to `sequential_hcl()` with default parameters chosen to generate more balanced heat colors than the basic `heat.colors()` function.
- `terrain_hcl()`: Convenience interface to `sequential_hcl()` with default parameters chosen to generate more balanced terrain colors than the basic `terrain.colors()` function.
- `diverging_hsv()`: Diverging palettes generated in HSV space rather than HCL space as in `diverging_hcl()`. This is provided for didactic purposes to contrast the more balanced HCL palettes with the more flashy and unbalanced HSV palettes.

Meanwhile, base R has also adopted the HCL-based palettes from **colorspace** into the function `hcl.colors()` in **grDevices** (Zeileis and Murrell 2019). This provides all the named palettes introduced in **colorspace** (with the same names, and defaulting to `"Viridis"`) but without the flexibility to modify or adapt existing palettes.

Moreover, the **grDevices** package in base R gained a new function `palette.colors()` (Zeileis, Murrell, Maechler, and Sarkar 2019) that provides various well-established qualitative color palettes that can not be approximated well by `qualitative_hcl()` due to pronounced variations in luminance and chroma. While a qualitative palette with fixed luminance and chroma is more balanced, a certain amount of variations in these properties might be necessary to make more colors distinguishable, especially for viewers with color vision deficiencies.

5. Palette visualization and assessment

The **colorspace** package provides several visualization functions for depicting one or more color palettes and their underlying properties. Color palettes can be visualized by:

- `swatchplot()`: Color swatches.
- `specplot()`: Spectrum of HCL and/or RGB trajectories.
- `hclplot()`: Trajectories in 2-dimensional HCL space projections.
- `demoplot()`: Illustrations of typical (and simplified) statistical graphics.

5.1. Color swatches

The function `swatchplot()` is a convenience function for displaying collections of palettes that can be specified as lists or matrices of hex color codes. Essentially, it is just a call to the base graphics `rect()` function but with heuristics for choosing default labels, margins, spacings, borders, etc. These heuristics are selected to work well for `hcl_palettes()`

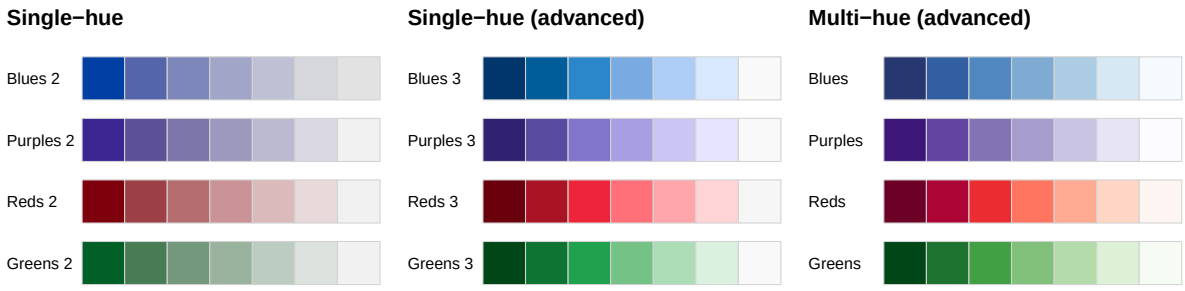

Figure 19: Variations of blue, purple, red, and green palettes with single hue and monotonic chroma (left), single hue and triangular chroma (center), and multiple hues and triangular chroma (right).

and might need further tweaking in future versions of the package. Thus, Figures 1–2 as well as Figures 10–14 all use `swatchplot()` internally. For a simple stand-alone illustration consider: `swatchplot("Palette" = sequential_hcl(5))`. Optionally, swatches emulating color vision deficiencies (see Section 6) can be added by setting `cvd = TRUE`.

Next, we demonstrate a more complex example of a `swatchplot()` with three matrices of sequential color palettes of blues, purples, reds, and greens (see Figure 19).

```
R> bprg <- c("Blues", "Purples", "Reds", "Greens")
R> swatchplot(
+   "Single-hue"            = t(sapply(paste(bprg, 2), sequential_hcl, n = 7)),
+   "Single-hue (advanced)" = t(sapply(paste(bprg, 3), sequential_hcl, n = 7)),
+   "Multi-hue (advanced)"  = t(sapply(bprg,           sequential_hcl, n = 7)),
+   nrow = 5, line = 5)
```

For all palettes, luminance increases monotonically to yield a proper sequential palette. However, the hue and chroma handling is somewhat different to emphasize different parts of the palette.

- *Single-hue:* In each palette the hue is fixed and chroma decreases monotonically (along with increasing luminance). This is typically sufficient to clearly bring out the extreme colors (dark/colorful vs. light gray).
- *Single-hue (advanced):* The hue is fixed (as above) but the chroma trajectory is triangular. Compared to the basic single-hue palette above, this better distinguishes the colors in the middle and not only the extremes.
- *Multi-hue (advanced):* As in the advanced single-hue palette, the chroma trajectory is triangular but additionally the hue varies slightly. This can further enhance the distinction of colors in the middle of the palette.

5.2. HCL (and RGB) spectrum

As the properties of a palette in terms of the perceptual dimensions *hue*, *chroma*, and *luminance* are not always clear from looking just at color swatches or (statistical) graphics based on these palettes, the `specplot()` function provides an explicit display for the coordinates of the HCL trajectory associated with a palette. This can bring out clearly various aspects,

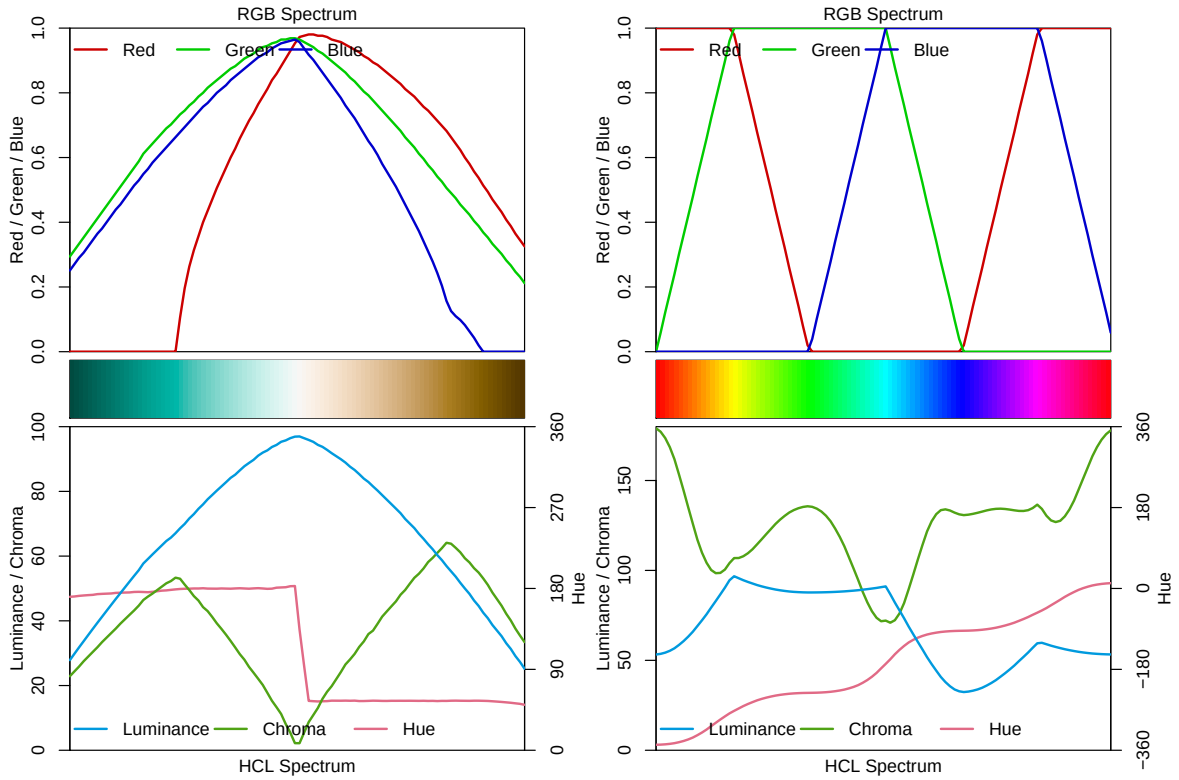

Figure 20: HCL spectrum of the balanced diverging `"Green-Brown"` palette (left panel) and the (in)famous and rather unbalanced `rainbow()` palette (right panel).

e.g., whether hue is constant, whether chroma is monotonic or triangular, and whether luminance is approximately constant (as in many qualitative palettes), monotonic (as in sequential palettes), or diverging.

The function first transforms a given color palette to its HCL (`polarLUV()`) coordinates. As the hues for low-chroma colors are not (or only poorly) identified, they are smoothed by default. Also, to avoid jumps from 0 to 360 or vice versa, the hue coordinates are shifted suitably. By default, the resulting trajectories in the HCL spectrum are visualized by a simple line plot where the x-axis gives the ordering of the colors in the palette. The y-axis depicts the following information:

- Hue is drawn in red and coordinates are indicated on the axis on the right with range $[0, 360]$ or (if necessary) $[-360, 360]$.
- Chroma is drawn in green with coordinates on the left axis. The range [0, 100] is used unless the palette necessitates higher chroma values.
- Luminance is drawn in blue with coordinates on the left axis in the range [0, 100].

Additionally, a color swatch for the palette is included. Optionally, a second spectrum for the corresponding trajectories of RGB coordinates can be included. However, this is usually just of interest for palettes created in RGB space (or simple transformations of RGB).

As spectrum plots have already been used for illustration in Figures 5 (for a qualitative palette) as well as Figures 6 and 17 (for sequential palettes), this section only provides a

couple of additional illustrations. The diverging `"Green-Brown"` palette is depicted in the left panel of Figure 20. It simply combines a green and a brown/yellow sequential single-hue palette, both with triangular chroma trajectory. Hue is constant in each "arm" of the palette and the chroma/luminance trajectories are rather balanced between both arms. In the center the palette passes through a light gray (with zero chroma) as the neutral value. By including the corresponding RGB spectrum in the top panel, it also becomes apparent that choosing such well-balanced palettes through trajectories in RGB color space is not straightforward. This balanced palette – based on relatively simple HCL trajectories – is contrasted with a poorly-balanced palette – based on simple linear RGB trajectories in the right panel of Figure 20. This depicts the RGB and HCL spectrum of the (in)famous RGB rainbow palette. (See Hawkins *et al.* 2014, for a plea why the RGB rainbow palette should be avoided in almost all scientific graphics.)

```
R> specplot(diverging_hcl(100, "Green-Brown"), rgb = TRUE)
R> specplot(rainbow(100), rgb = TRUE)
```

The RGB spectrum of the rainbow palette shows that the trajectories are quite simple in RGB space but lead to substantial variations in chroma and (more importantly) luminance. This is why this palette is not suitable for encoding underlying data in statistical graphics. See also the related discussion of color vision deficiency in Section 6.

5.3. Trajectories in HCL space

While the `specplot()` function above works well for bringing out the HCL coordinates associated with a given palette, it does not show how the palette fits into the HCL space. For example, it is not so clear whether high chroma values are close to the maximum possible for a given hue. Thus, it cannot be easily judged how the parameters of the hue, chroma, and luminance trajectories can be modified to obtain another palette.

Therefore, the `hclplot()` is another visualization of the HCL coordinates associated with a palette. It does so by collapsing over one of the coordinates (either the hue H or the luminance L) and displaying a heatmap of colors combining the remaining two dimensions. The coordinates for the given color palette are highlighted to bring out its trajectory. In case the hue is really fixed (as in single-hue sequential palettes) or the luminance is really fixed (as in the qualitative palettes), collapsing is straightforward. However, when the coordinate that is collapsed over is not actually constant in the palette, a simple bivariate linear model is used to capture how the collapsed coordinate varies along with the two displayed coordinates.

The function `hclplot()` has been designed to work well with the `hcl_palettes()` in this package. While it is possible to apply it to other color palettes as well, the results might look weird or confusing if these palettes are constructed very differently (e.g., like the highly saturated base R palettes). To infer the default `type` of projection, `hclplot()` assesses the luminance trajectory and sets the default correspondingly:

- `type = "qualitative"` if luminance is approximately constant.
- `type = "sequential"` if luminance is monotonic.
- `type = "diverging"` if luminance is diverging with two monotonic "arms" in the trajectory.

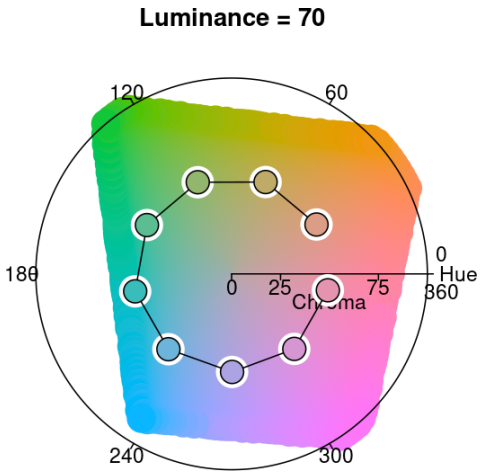

Figure 21: Hue-chroma plane with luminance fixed at $L = 70$ along with the qualitative `"Dynamic"` palette with varying hue H and chroma fixed at $C = 50$.

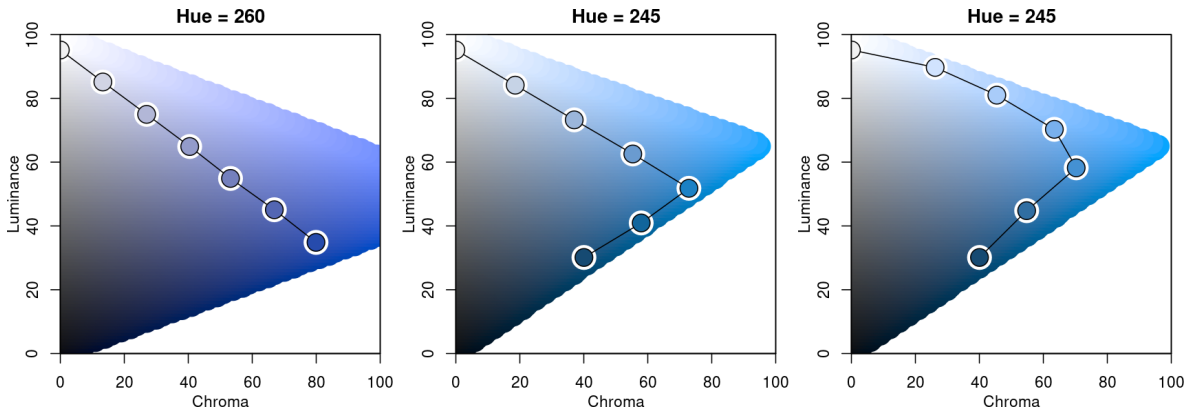

Figure 22: Luminance-chroma planes with variations of blue sequential single-hue palettes (similar to `"Blues 2"` and `"Blues 3"`). Left: Linear chroma for $H = 260$. Center: Triangular chroma for $H = 245$. Right: Power-transformed triangular chroma for $H = 245$.

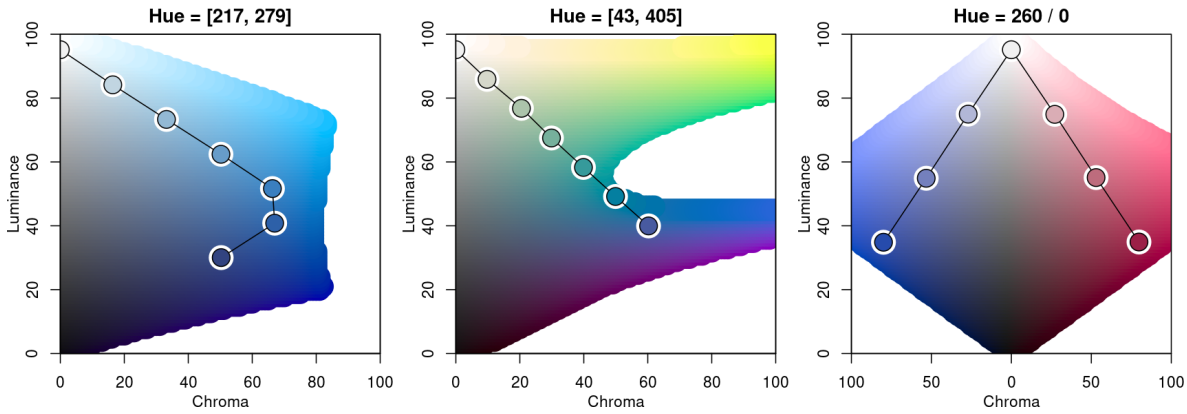

Figure 23: Luminance-chroma planes with blue multi-hue palette and triangular chroma (left), blue-yellow multi-hue palette and linear chroma (center), and diverging blue-red palette with balanced linear chroma.

Thus, for qualitative palettes – where luminance and chroma are fixed – the varying hue is displayed in a projection onto the hue-chroma plane at a given fixed luminance (Figure 21):

```
R> hclplot(qualitative_hcl(9, "Dynamic"))
```

Figure 22 compares three single-hue sequential palettes by projection to the luminance-chroma plane for the given fixed hue. In the left panel the hue 260 is used with a simple linear chroma trajectory. The other two panels employ a triangular chroma trajectory for hue 245, either with a piecewise-linear (center) or power-transformed (right) trajectory.

```
R> par(mfrow = c(1, 3))
R> hclplot(sequential_hcl(7, h = 260, c = 80, l = c(35, 95), power = 1))
R> hclplot(sequential_hcl(7, h = 245, c = c(40, 75, 0), l = c(30, 95),
+    power = 1))
R> hclplot(sequential_hcl(7, h = 245, c = c(40, 75, 0), l = c(30, 95),
+    power = c(0.8, 1.4)))
```

Note that for $H = 260$ it is possible to go to dark colors (low luminance) with high chroma while this is not possible to the same extent for $H = 245$ due to the distorted shape of the HCL space. Hence, chroma has to be decreased when proceeding to the dark low-luminance colors.

Finally, Figure 23 compares two multi-hue sequential palettes along with a diverging palette.

```
R> par(mfrow = c(1, 3))
R> hclplot(sequential_hcl(7, h = c(260, 220), c = c(50, 75, 0),
+    l = c(30, 95), power = 1))
R> hclplot(sequential_hcl(7, h = c(260, 60), c = 60, l = c(40, 95),
+    power = 1))
R> hclplot(diverging_hcl(7, h = c(260, 0), c = 80, l = c(35, 95),
+    power = 1))
```

The multi-hue palette on the left employs a small hue range, resulting in a palette of "blues" just with slightly more distinction of the middle colors in the palette. In contrast, the multi-hue "blue-yellow" palette in the center panel uses a large hue range, resulting in more color contrasts throughout the palette. Finally, the balanced diverging palette in the right panel is constructed from two simple single-hue sequential palettes (for hues 260/blue and 0/red) that are completely balanced between the two "arms" of the palette.

5.4. Demonstration of statistical graphics

To demonstrate how different kinds of color palettes work in different kinds of statistical displays, `demoplot()` provides a simple convenience interface to some base graphics with (mostly artificial) data sets. As a first overview, Figure 24 displays all built-in demos with the same sequential heat colors palette: `sequential_hcl(5, "Heat")`. All types of demos can, in principle, deal with arbitrarily many colors from any palette, but the graphics differ in various respects such as:

- Working best for fewer colors (e.g., bar, pie, scatter, lines, ...) vs. many colors (e.g., heatmap, perspective, ...).

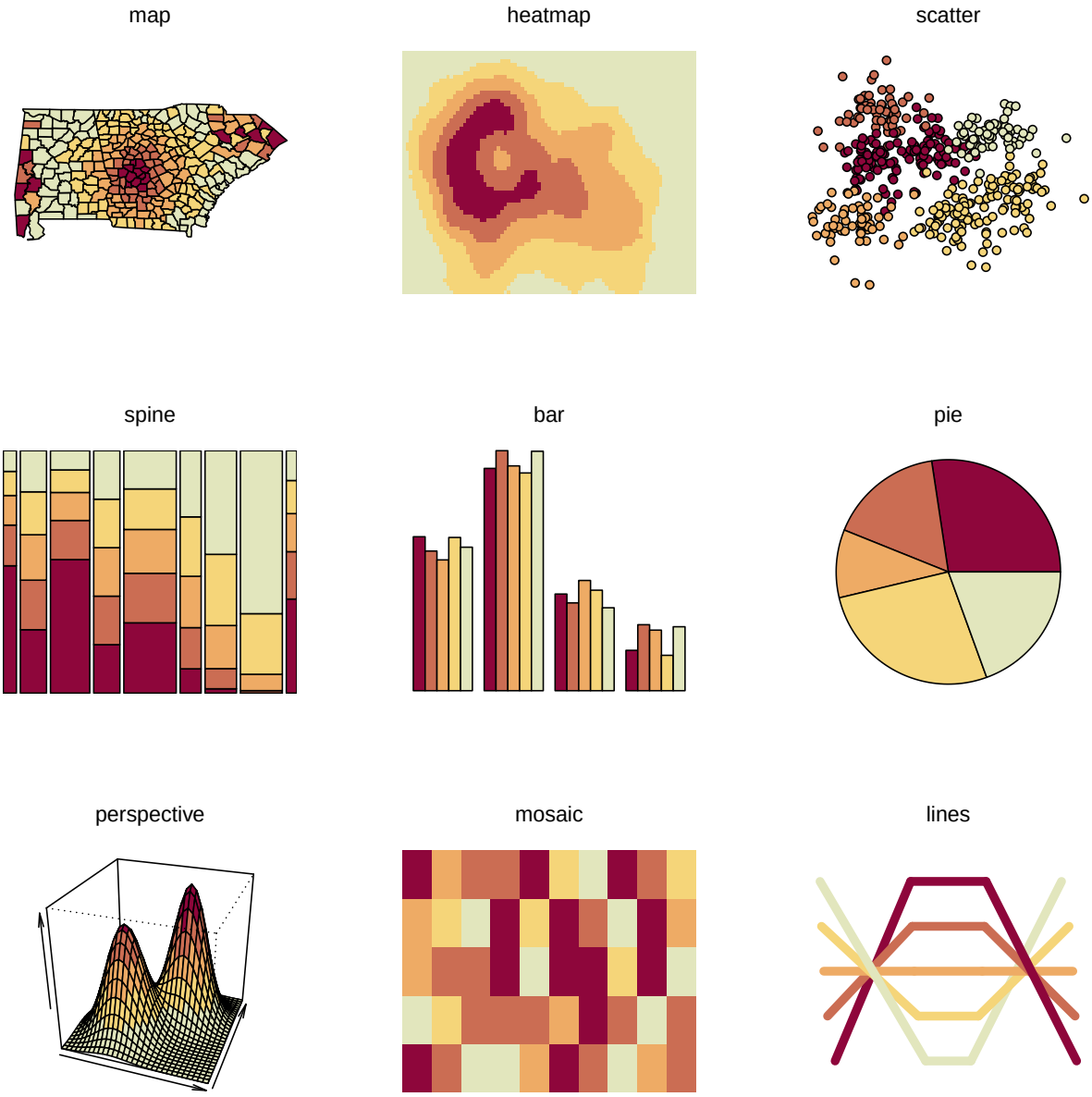

Figure 24: All built-in `demoplot` types with the same `sequential_hcl(5, "Heat")` palette.

- Intended for categorical data (e.g., bar, pie, ...) vs. continuous numeric data (e.g., heatmap, perspective, ...).
- Shading areas (e.g., map, bar, pie, ...) vs. coloring points or lines (scatter, lines).

Hence, in the following Figures 25–27 some further illustrations are organized by type of palette, using suitable demos for the particular palettes.

Qualitative palettes: Light pastel colors typically work better for shading areas (pie, left) while darker and more colorful palettes are usually preferred for points (center) or lines (right).

```
R> par(mfrow = c(1, 3))
R> demoplot(qualitative_hcl(4, "Pastel 1"), type = "pie")
```

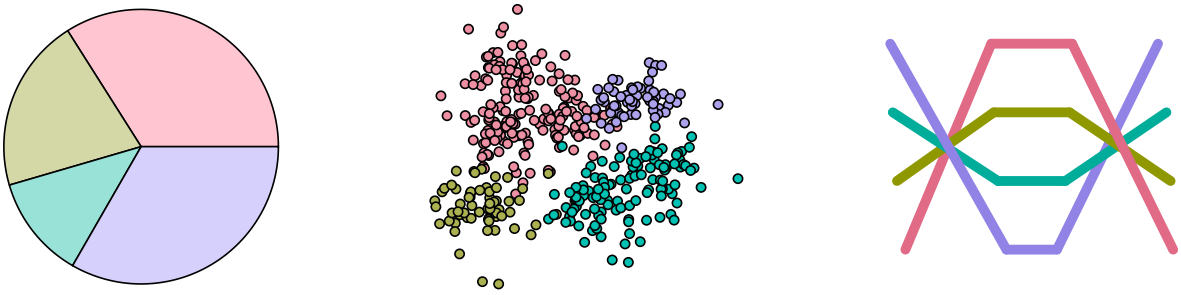

Figure 25: Examples for `demoplot()` with different `qualitative_hcl()` palettes.

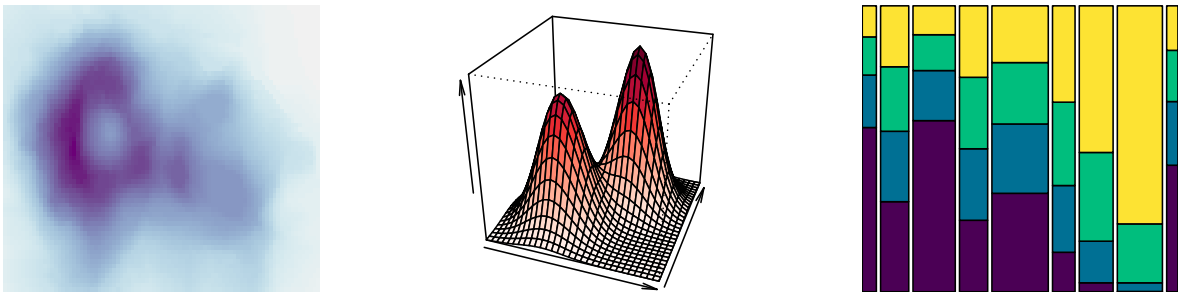

Figure 26: Examples for `demoplot()` with different `sequential_hcl()` palettes.

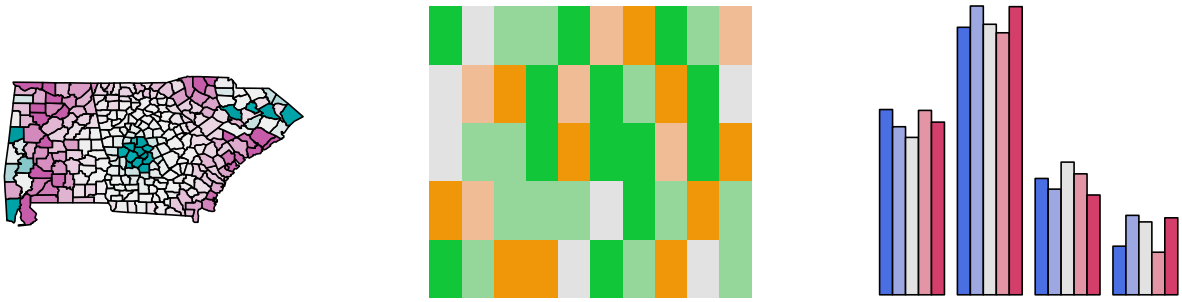

Figure 27: Examples for `demoplot()` with different `diverging_hcl()` palettes.

```
R> demoplot(qualitative_hcl(4, "Set 2"), type = "scatter")
R> demoplot(qualitative_hcl(4, "Dark 3"), type = "lines")
```

Sequential palettes: Heatmaps (left) or perspective plots (center) often employ almost continuous gradients with strong luminance contrasts. In contrast, when only a few ordered categories are to be displayed (e.g., in a spine plot, right) more colorful sequential palettes like the viridis palette can be useful.

```
R> par(mfrow = c(1, 3))
R> demoplot(sequential_hcl(99, "Purple-Blue"), type = "heatmap")
R> demoplot(sequential_hcl(99, "Reds"), type = "perspective")
R> demoplot(sequential_hcl( 4, "Viridis"), type = "spine")
```

Diverging palettes: In some displays (such as the map, left), it is useful to employ an almost continuous gradient with strong luminance contrast to bring out the extremes. Here, this

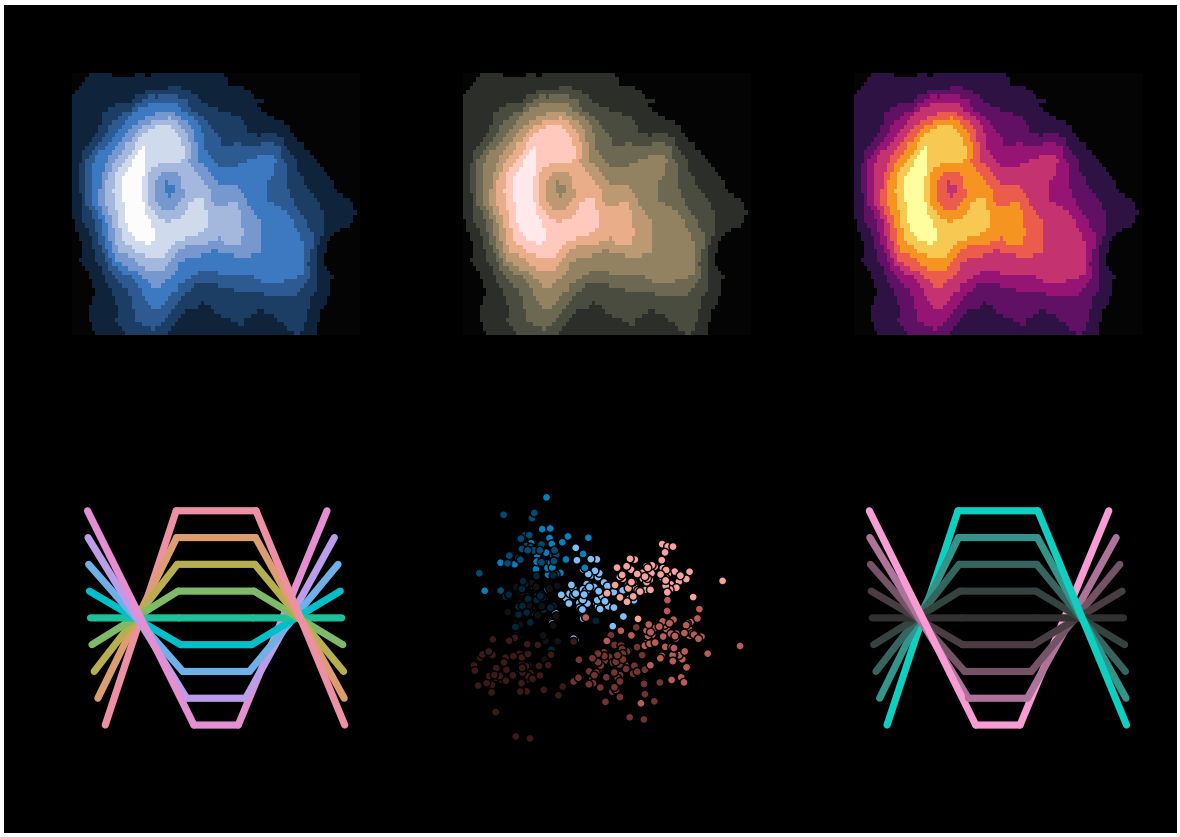

Figure 28: Examples for `demoplot()` with different palettes that work well on a black/dark background.

contrast is amplified by a larger power transformation emphasizing the extremes even further. In contrast, when fewer colors are needed more colorful palettes with lower luminance contrasts can be desired. This is exemplified by a mosaic (center) and bar plot (right).

```
R> par(mfrow = c(1, 3))
R> demoplot(diverging_hcl(99, "Tropic", power = 2.5), type = "map")
R> demoplot(diverging_hcl( 5, "Green-Orange"), type = "mosaic")
R> demoplot(diverging_hcl( 5, "Blue-Red 2"), type = "bar")
```

Figures 25–27 focus on palettes designed for light/white backgrounds. Therefore, to conclude, some palettes are highlighted in Figure 28 that work well on dark/black backgrounds.

```
R> par(mfrow = c(2, 3), bg = "black")
R> demoplot(sequential_hcl(9, "Oslo"), "heatmap")
R> demoplot(sequential_hcl(9, "Turku"), "heatmap")
R> demoplot(sequential_hcl(9, "Inferno", rev = TRUE), "heatmap")
R> demoplot(qualitative_hcl(9, "Set 2"), "lines")
R> demoplot(diverging_hcl(9, "Berlin"), "scatter")
R> demoplot(diverging_hcl(9, "Cyan-Magenta", l2 = 20), "lines")
```

6. Color vision deficiency emulation

Different kinds of limitations can be emulated using the physiologically-based model for simulating color vision deficiency (CVD) of Machado, Oliveira, and Fernandes (2009): deuteranomaly (green cone cells defective), protanomaly (red cone cells defective), and tritanomaly (blue cone cells defective). While most other CVD simulations handle only dichromacy, where one of three cones is non-functional, Machado *et al.* (2009) provide a unified model of both dichromacy and anomalous trichromacy, where one cone has shifted spectral sensitivity. As anomalous trichromacy is the most common form of color vision deficiency, it is important to emulate along with the rarer, but more severe dichromacy. Below we briefly describe our R interface to these emulation techniques and show them in practice for a heatmap with sequential palette. Another example with a diverging palette is available at `http://colorspace.R-Forge.R-project.org/articles/color_vision_deficiency.html`. Finally, CVD emulation is particularly useful for bringing out why the RGB rainbow palette is almost always a bad choice in scientific displays. See `http://colorspace.R-Forge.R-project.org/articles/endrainbow.html` for further illustrations.

6.1. R functions

The workhorse function to emulate color vision deficiencies is `simulate_cvd()` which can take any vector of valid R colors and transform them according to a certain CVD transformation matrix and transformation equation. The transformation matrices have been established by Machado *et al.* (2009) and are provided in objects `protanomaly_cvd`, `deutanomaly_cvd`, and `tritanomaly_cvd`. The convenience interfaces `deutan()`, `protan()`, and `tritan()` are the high-level functions for simulating the corresponding kind of color blindness with a given `severity` (calling `simulate_cvd()` internally). A severity of `1` corresponds to dichromacy, `0` to normal color vision, and intermediate values to varying severities of anomalous trichromacy. For further guidance on color blindness in relation to statistical graphics see Lumley (2006) which accompanies the R package **dichromat** (Lumley 2013) and is based on earlier emulation techniques (Viénot, Brettel, Ott, M'Barek, and Mollon 1995; Brettel, Viénot, and Mollon 1997; Viénot, Brettel, and Mollon 1999).

6.2. Illustration: Heatmap with sequential palette

To illustrate that poor color choices can severely reduce the usefulness of a statistical graphic for readers with color vision deficiencies, we employ the infamous RGB rainbow color palette

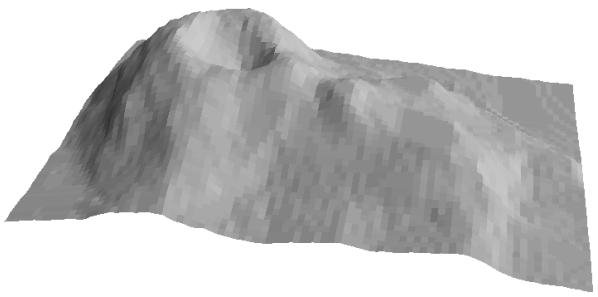

Figure 29: Perspective visualization of Maunga Whau `volcano` data (Mount Eden, Auckland, New Zealand).

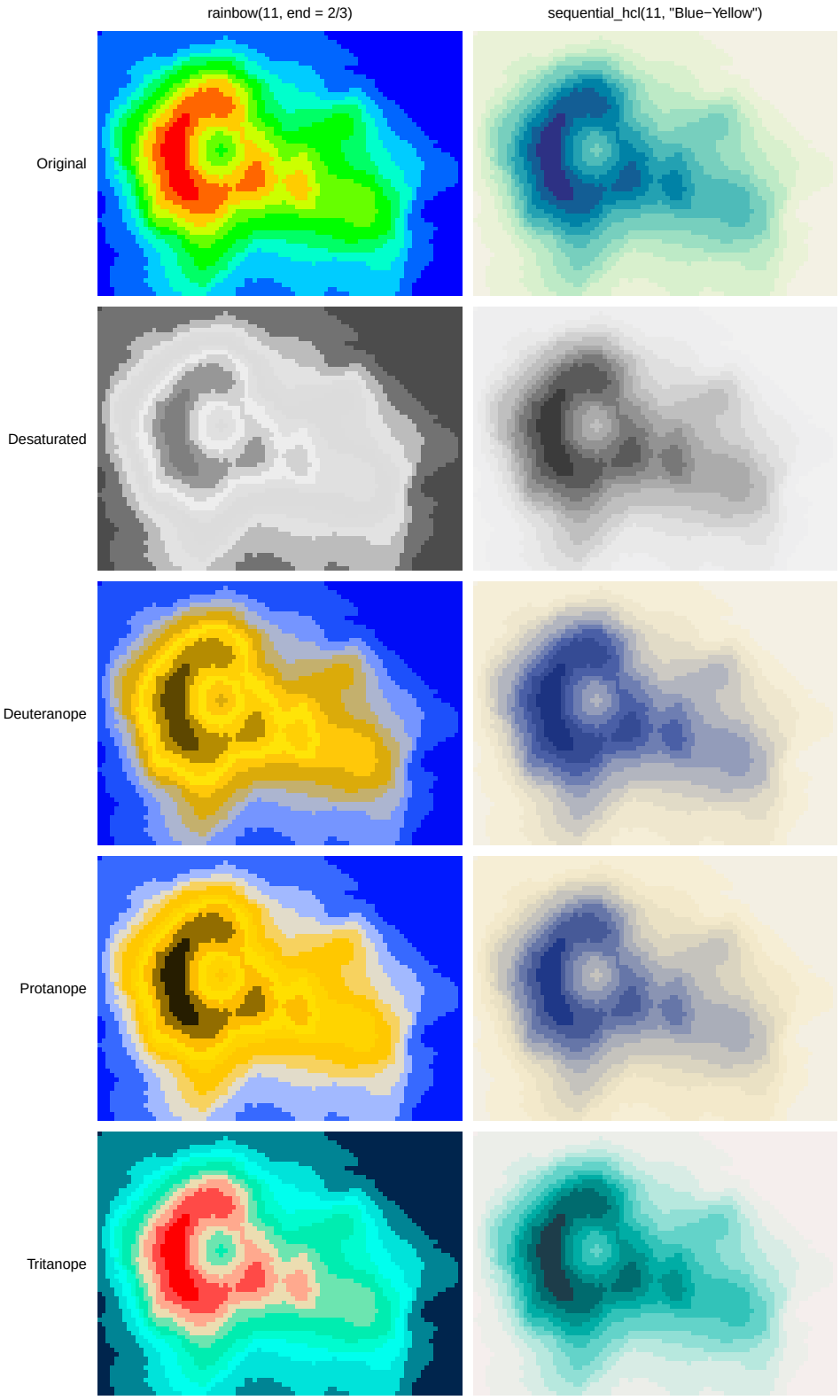

Figure 30: Heatmap of Maunga Whau `volcano` data with RGB rainbow (left) and HCL-based blue-yellow palette (right). The first row shows the original color palettes while subsequent rows emulate various color deficiencies.

in a heatmap. In base R this can be generated by `rainbow(11, end = 2/3)` ranging from red (for high values) to blue (for low values). The poor results for the RGB rainbow palette are contrasted in Figure 30 with a proper sequential palette ranging from dark blue to light yellow: `sequential_hcl(11, "Blue-Yellow")`.

The statistical graphic employed for illustration is a heatmap of the well-known Maunga Whau `volcano` data from base R. This heatmap is easily available as `demoplot(x, "heatmap")` where `x` is the color vector to be used, e.g.,

```
R> rainbow(11, end = 2/3)

 [1] "#FF0000FF" "#FF6600FF" "#FFCC00FF" "#CCFF00FF" "#66FF00FF"
 [6] "#00FF00FF" "#00FF66FF" "#00FFCCFF" "#00CCFFFF" "#0066FFFF"
[11] "#0000FFFF"

R> deutan(rainbow(11, end = 2/3))

 [1] "#5D4700FF" "#B58C01FF" "#FFD005FF" "#FFE408FF" "#FFC809FF"
 [6] "#DBAB0AFF" "#C4B06DFF" "#ACB5D0FF" "#7595FFFF" "#1D50FBFF"
[11] "#000CF7FF"
```

and so on. To aid the interpretation of the heatmap a perspective display using only gray shades is provided in Figure 29, providing another intuitive display of what the terrain around Maunga Whau looks like.

Subsequently, all combinations of palette and color vision deficiency are visualized. Additionally, a grayscale version is created with `desaturate()`. This clearly shows how poorly the RGB rainbow performs, often giving quite misleading impressions of the terrain around Maunga Whau. In contrast, the HCL-based blue-yellow palette works reasonably well in all settings. The most important problem of the RGB rainbow is that it is not monotonic in luminance, making correct interpretation quite hard. Moreover, the red-green contrasts deteriorate substantially in the dichromatic emulations.

7. Apps for choosing colors and palettes interactively

To facilitate exploring the package and employing it when working with colors, several graphical user interfaces (GUIs) are provided within the package as **shiny** apps (Chang *et al.* 2020). All of these GUIs/apps can be run locally from within R and are also provided online at http://hclwizard.org/.

- *Palette constructor:* `choose_palette()` or `hclwizard()` or `hcl_wizard()`.
- *Color picker:* `choose_color()` or equivalently `hcl_color_picker()`.
- *Color vision deficiency emulator:* `cvd_emulator()`.

In addition to the **shiny** version, the *palette constructor* app is also available as a Tcl/Tk GUI via R package **tcltk** shipped with base R (R Core Team 2020). The **tcltk** version can only be run locally and is considerably faster while the **shiny** version has a nicer interface with more features and can be run online. The `choose_palette()` function by default starts the **tcltk** version while `hclwizard()`/`hcl_wizard()` by default start the **shiny** version.

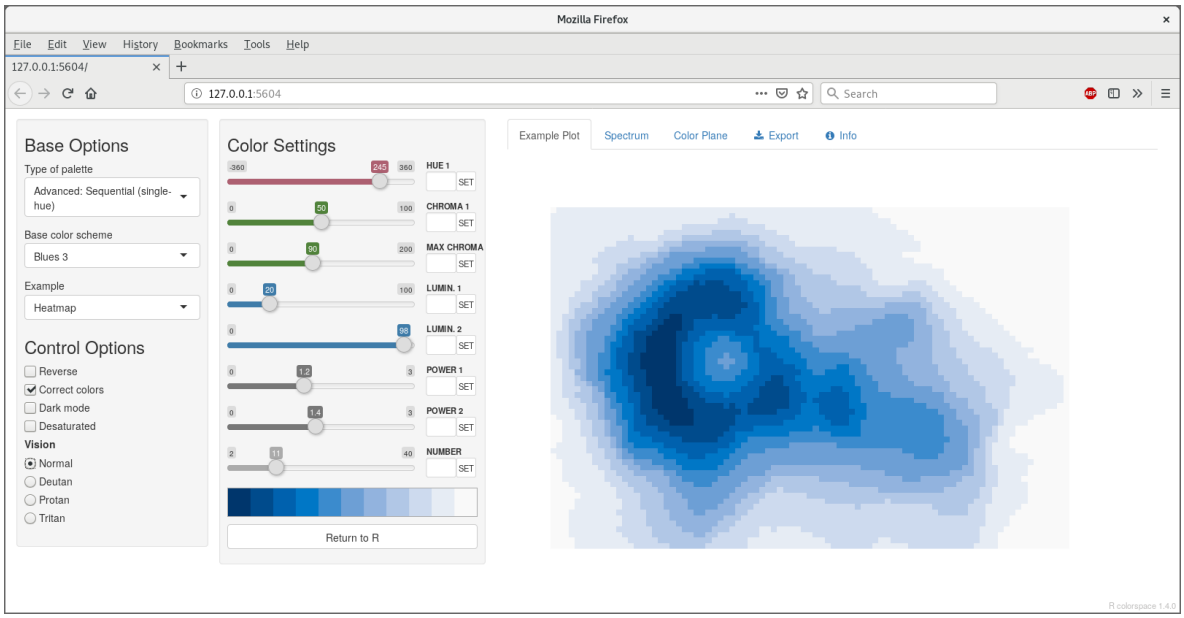

Figure 31: App for interactively choosing HCL-based color palettes: `choose_color()`/`hclwizard()`.

7.1. Choose palettes with the HCL color model

The **shiny** version of the palette constructor GUI is shown in Figure 31. It interfaces the `qualitative_hcl()`, `sequential_hcl()`, and `diverging_hcl()` palettes from Section 4. The GUIs allow for interactive modification of the arguments of the respective palette-generating functions, i.e., starting/ending hue, minimal/maximal chroma, minimal/maximal luminance, and power transformations that control how quickly/slowly chroma and/or luminance are changed through the palette. Subsets of the parameters may not be applicable depending on the type of palette chosen.

Optionally, the active palette can be illustrated by using a `specplot()` (see Section 5.2), `hclplot()` (see Section 5.3), or `demoplot()` (see Section 5.4), and assessed using emulation of color vision deficiencies (see Section 6). To facilitate generation of palettes for black/dark backgrounds, a "dark mode" of the GUIs is also available.

The app has been influenced considerably by **ColorBrewer.org** (Harrower and Brewer 2003). Similarities include the selection of a qualitative, sequential, or diverging palette from a list of predefined colors along with an example visualization. However, unlike **ColorBrewer.org** our **shiny** app allows tweaking the HCL parameters underlying each palette. This makes the app much more flexible but also more complex, potentially requiring more thought and experience. Due to the flexibility, our app cannot automatically judge safety regarding color vision deficiencies and printers/photocopiers (as **ColorBrewer.org** does) but instead it allows emulation of color vision deficiencies and desaturation. Finally, **ColorBrewer.org** is geared towards cartography (albeit its palettes are useful much more generally) while our **shiny** app includes a broader range of illustrative displays.

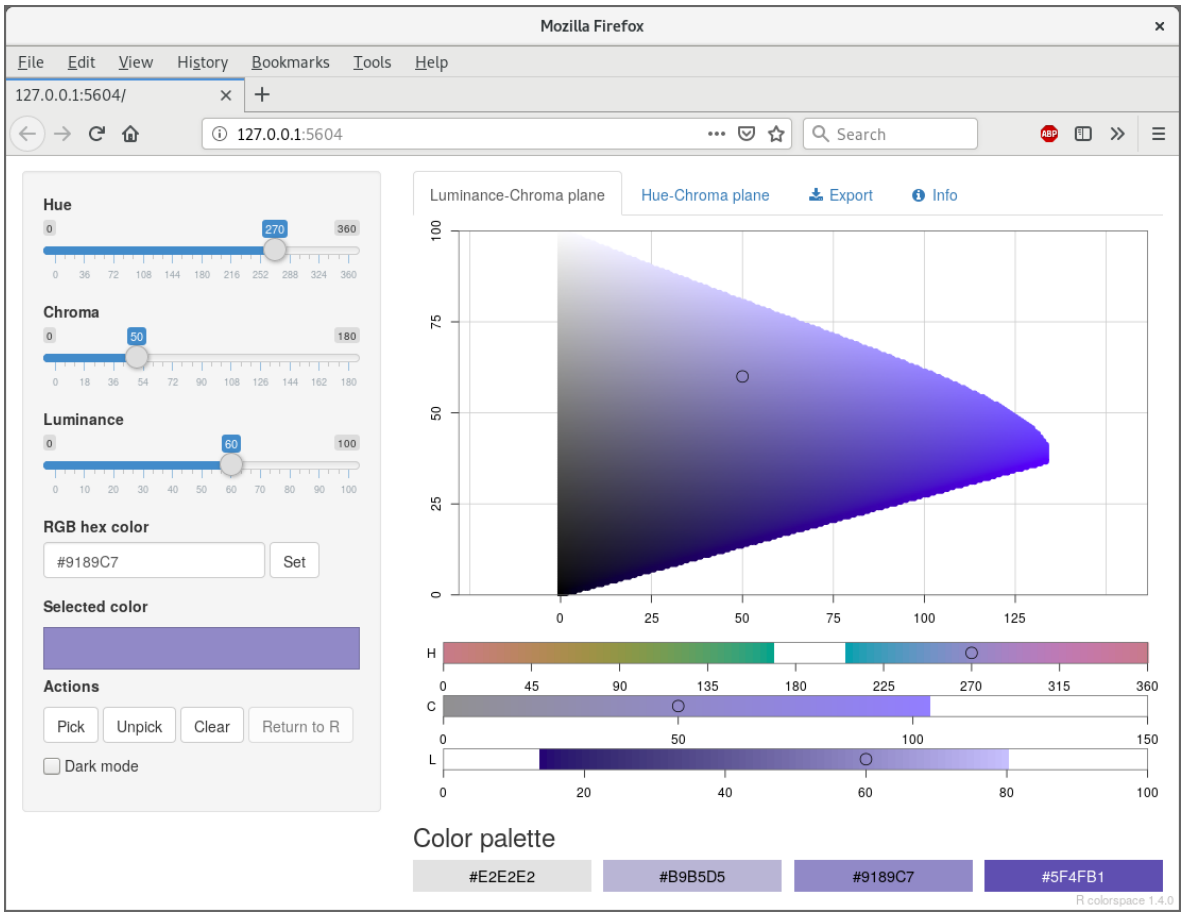

Figure 32: App for interactively choosing individual colors in HCL space: `choose_color()`/`hcl_color_picker()`.

7.2. Choose individual colors with the HCL color model

This GUI can be started with either `choose_color()` or equivalently `hcl_color_picker()`. It shows the HCL color space either as a hue-chroma plane for a given luminance value or as a luminance-chroma plane for a given hue. Colors can be entered by:

- Clicking on a color coordinate in the hue-chroma or luminance-chroma plane.
- Specifying the hue/chroma/luminance values via sliders.
- Entering an RGB hex code.

By repeating the selection a palette of colors can be constructed and returned within R for subsequent usage in visualizations.

7.3. Emulate color vision deficiencies

This GUI can be started with `cvd_emulator()`. It supports uploading a raster image in JPG or PNG format which is then checked for various kinds of color vision deficiencies at the selected severity. By default the severity is set to 100% and all supported kinds of color vision deficiency are checked for.

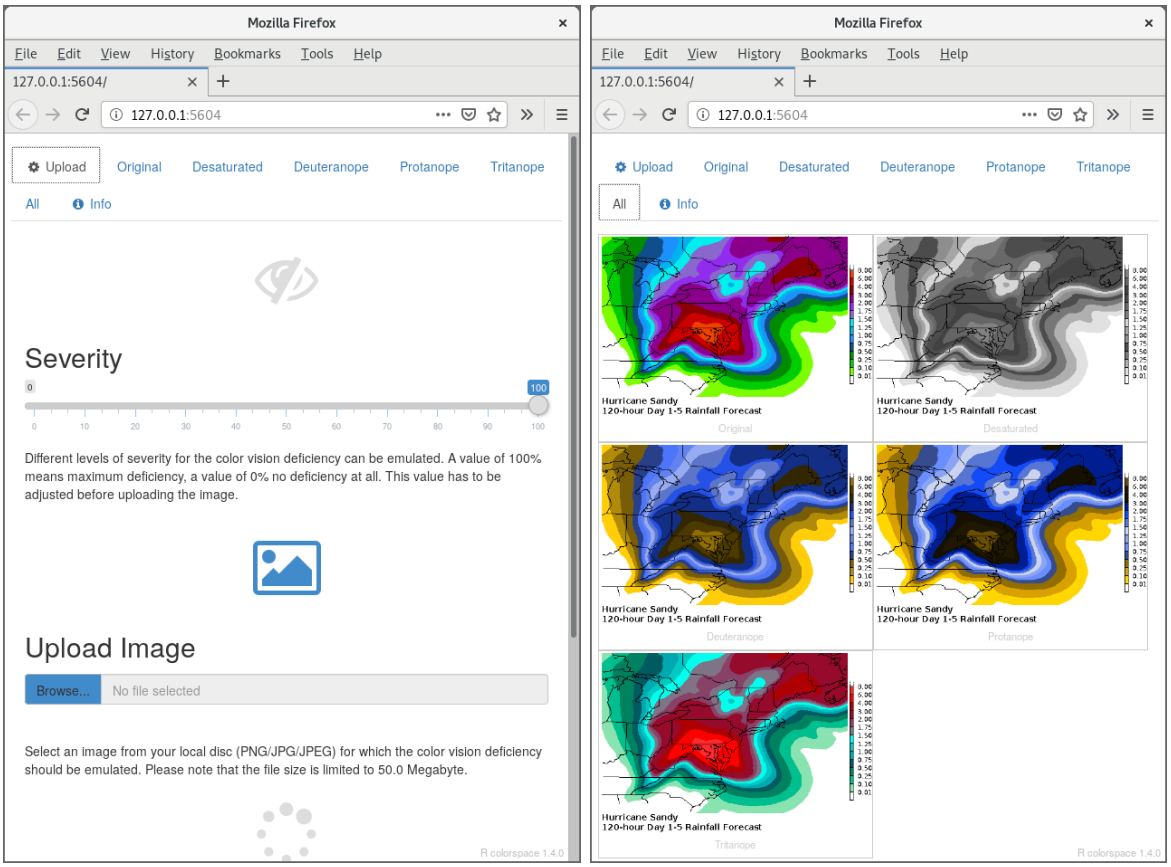

Figure 33: App for emulating color vision deficiencies for uploaded raster images: `cvd_emulator()`.

8. Color manipulation and utilities

The **colorspace** package provides several color manipulation utilities that are useful for creating, assessing, or transforming color palettes, namely:

- `desaturate()`: Desaturate colors by chroma removal in HCL space.
- `darken()` and `lighten()`: Algorithmically lighten or darken colors in HCL and/or HLS space.
- `max_chroma()`: Compute maximum chroma for given hue and luminance in HCL space.
- `mixcolor()`: Additively mix two colors by computing their convex combination.

8.1. Desaturation in HCL space

Desaturation should map a given color to the gray with the same "brightness". In principle, any perceptually-based color model (HCL, HLS, HSV, ...) could be employed for this but HCL works particularly well because its coordinates capture the perceptual properties better than most other color models.

The `desaturate()` function converts any given hex color code or named R color to the corresponding HCL coordinates and sets the chroma to zero. Thus, only the luminance matters

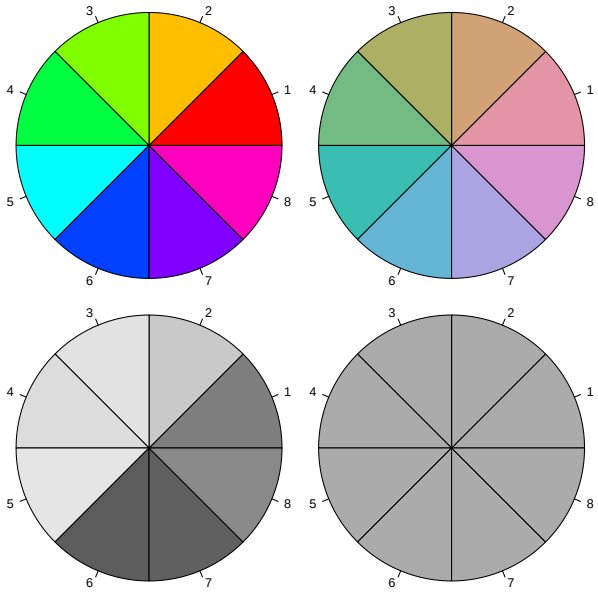

Figure 34: Color wheels in RGB (left) and HCL (right) space in color (top) and desaturated grayscale (bottom).

which captures the "brightness" mentioned above. Finally, the resulting HCL coordinates are transformed back to hex color codes for use in R. First, `desaturate()` is used to desaturate a vector of R color names:

```
R> desaturate(c("white", "orange", "blue", "black"))

[1] "#FFFFFF" "#B8B8B8" "#4C4C4C" "#000000"
```

Notice that the hex codes corresponding to three coordinates in sRGB space are always the same, thus corresponding to gray colors (due to the same amount of red, green, and blue). Analogously, hex color codes can also be transformed – in this case RGB rainbow colors from the base R function `rainbow()`:

```
R> rainbow(3)

[1] "#FF0000FF" "#00FF00FF" "#0000FFFF"

R> desaturate(rainbow(3))

[1] "#7F7F7FFF" "#DCDCDCFF" "#4C4C4CFF"
```

Even this simple example suffices to show that the three RGB rainbow colors have very different grayscale levels. This deficiency is even clearer when using a full color wheel (of colors with hues in [0, 360] degrees). While the RGB `rainbow()` is very unbalanced, the HCL `rainbow_hcl()` (or also `qualitative_hcl()`) is (by design) balanced with respect to luminance.

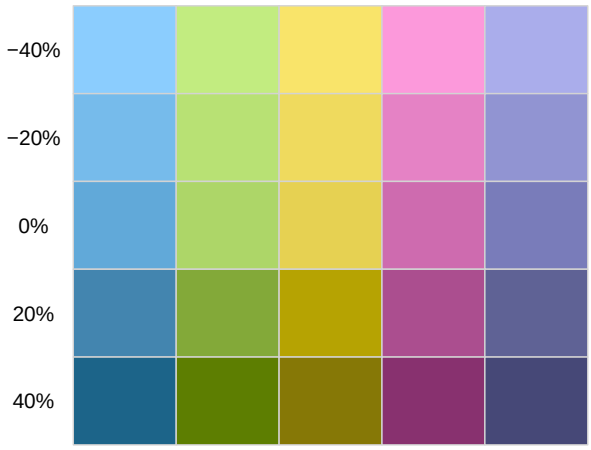

Figure 35: Okabe-Ito palette (0%) along with two levels of both lightening and darkening, respectively.

```
R> wheel <- function(col, radius = 1, ...)
+    pie(rep(1, length(col)), col = col, radius = radius, ...)
R> par(mar = rep(0.5, 4), mfrow = c(2, 2))
R> wheel(rainbow(8))
R> wheel(rainbow_hcl(8))
R> wheel(desaturate(rainbow(8)))
R> wheel(desaturate(rainbow_hcl(8)))
```

8.2. Lighten or darken colors

In principle, a similar approach for lightening and darkening colors can be employed as for desaturation above. The colors can simply be transformed to HCL space and then the luminance can either be decreased (turning the color darker) or increased (turning it lighter) while preserving the hue and chroma coordinates. This strategy typically works well for lightening colors, although in some situations the result can be somewhat too colorful. Conversely, when darkening rather light colors with little chroma, this can result in rather gray colors.

In these situations, an alternative might be to apply the analogous strategy in HLS space which is frequently used in HTML style sheets. However, this strategy may also yield colors that are either too gray or too colorful. A compromise that sometimes works well is to adjust the luminance coordinate in HCL space but to take the chroma coordinate corresponding to the HLS transformation.

We have found that typically the HCL-based transformation performs best for lightening colors and this is hence the default in `lighten()`. For darkening colors, the combined strategy often works best and is hence the default in `darken()`. In either case it is recommended to try the other available strategies in case the default yields unexpected results.

Regardless of the chosen color space, the adjustment of the `L` component by a certain `amount` can occur by two methods, relative (the default) or absolute. For example, `L - 100 * amount` is used for absolute darkening, or `L * (1 - amount)` for relative darkening. See `?lighten` and `?darken` for more details.

For illustration the qualitative palette suggested by Okabe and Ito (2008) is transformed by two levels of both lightening and darkening, respectively (see Figure 35).

```
R> oi <- c("#61A9D9", "#ADD668", "#E6D152", "#CE6BAF", "#797CBA")
R> swatchplot( "-40%" = lighten(oi, 0.4), "-20%" = lighten(oi, 0.2),
+    "  0%" = oi, " 20%" = darken(oi, 0.2), " 40%" = darken(oi, 0.4),
+    off = c(0, 0))
```

8.3. Adjust transparency of colors

Alpha transparency is useful for making colors semi-transparent, e.g., for overlaying different elements in graphics (Wikipedia 2020i). An alpha value (or alpha channel) of 0 (or 00 in hex strings) corresponds to fully transparent and an alpha value of 1 (or FF in hex strings) corresponds to fully opaque. If a color hex string in R does not provide an explicit alpha transparency, the color is assumed to be fully opaque.

The `adjust_transparency()` function can be used to adjust the alpha transparency of a set of colors. It always returns a hex color specification. This hex color can have the alpha transparency added/removed/modified depending on the specification of the argument `alpha`:

- `alpha = NULL`: Returns a hex vector with alpha transparency only if needed. Thus, it keeps the alpha transparency for the colors (if any) but only if different from opaque.
- `alpha = TRUE`: Returns a hex vector with alpha transparency for all colors, using opaque (FF) as the default if missing.
- `alpha = FALSE`: Returns a hex vector without alpha transparency for all colors (even if the original colors had non-opaque alpha).
- `alpha` numeric: Returns a hex vector with alpha transparency for all colors set to the `alpha` argument (recycled if necessary).

For illustration, the transparency of a single black color is modified to three alpha levels: fully transparent, semi-transparent, and fully opaque, respectively. Black can be equivalently specified by name (`"black"`), hex string (`"#000000"`), or integer position in the palette (`1`).

```
R> adjust_transparency("black",   alpha = c(0, 0.5, 1))

[1] "#00000000" "#00000080" "#000000FF"

R> adjust_transparency("#000000", alpha = c(0, 0.5, 1))

[1] "#00000000" "#00000080" "#000000FF"

R> adjust_transparency(1,         alpha = c(0, 0.5, 1))

[1] "#00000000" "#00000080" "#000000FF"
```

Subsequently, different settings of `alpha` are illustrated for adjusting a vector with three shades of gray, specified by name (`gray`, opaque), opaque hex string (`"#BEBEBE"`), and semi-transparent hex string (`"#BEBEBE80"`). Four types of adjustment are shown: only if necessary (`alpha = NULL`), add (`alpha = TRUE`), remove (`alpha = FALSE`), or modify (`alpha = 0.8`).

```
R> x <- c("gray", "#BEBEBE", "#BEBEBE80")
R> adjust_transparency(x, alpha = NULL)

[1] "#BEBEBE"   "#BEBEBE"   "#BEBEBE80"

R> adjust_transparency(x, alpha = TRUE)

[1] "#BEBEBEFF" "#BEBEBEFF" "#BEBEBE80"

R> adjust_transparency(x, alpha = FALSE)

[1] "#BEBEBE" "#BEBEBE" "#BEBEBE"

R> adjust_transparency(x, alpha = 0.8)

[1] "#BEBEBECC" "#BEBEBECC" "#BEBEBECC"
```

8.4. Maximum chroma for given hue and luminance

As the possible combinations of chroma and luminance in HCL space depend on hue, it is not obvious which trajectories through HCL space are possible prior to trying a specific HCL coordinate by calling `polarLUV()`. To avoid having to fix up the color upon conversion to RGB `hex()` color codes, the `max_chroma()` function computes (approximately) the maximum chroma possible. For illustration we show that for given luminance (here: $L = 50$) the maximum chroma varies substantially with hue:

```
R> max_chroma(h = seq(0, 360, by = 60), l = 50)

[1] 137.96  59.99  69.06  39.81  65.45 119.54 137.96
```

Similarly, maximum chroma also varies substantially across luminance values for a given hue (here: $H = 120$, green):

```
R> max_chroma(h = 120, l = seq(0, 100, by = 20))

[1]   0.00  28.04  55.35  82.79 110.28   0.00
```

8.5. Additive mixing of two colors

In additive color models like `RGB()` or `XYZ()` it can be useful to combine colors by additive mixing. Below a fully saturated red and green are mixed, yielding a medium brownish yellow.

```
R> R <- RGB(1, 0, 0)
R> G <- RGB(0, 1, 0)
R> Y <- mixcolor(0.5, R, G)
R> Y

       R   G B
[1,] 0.5 0.5 0
```

9. Summary and discussion

This paper provides an overview of the broad capabilities of the **colorspace** package for selecting individual colors or color palettes, manipulating these colors, and employing them in various kinds of visualizations.

In particular, the package provides various qualitative, sequential, and diverging palettes derived by relatively simple trajectories in HCL (hue-chroma-luminance) space. In contrast to many other packages providing modern balanced color palettes (such as **ColorBrewer.org**, **CARTO**, **viridis**, or **scico**) special emphasis is given to flexibility of the palettes, which can be adjusted to the particular needs of a given data visualization. The paper also provides various tips and tricks for choosing an effective palette in a given situation. Further useful guidance is provided in many sources, including: Ware (1988), Okabe and Ito (2008), Aigner (2010), Stauffer *et al.* (2015), Zhang (2015), Rost (2018), Wilke (2019), and Ciechanowski (2019), among many others.

There are other R packages that can complement the palettes provided by **colorspace**. **Polychrome** (Coombes and Brock 2020; Coombes, Brock, Abrams, and Abruzzo 2019) implements strategies for qualitative palettes with many "categories". While the qualitative palettes in Section 4 yield only about 6–8 clearly distinguishable colors due to the fixed chroma and luminance, **Polychrome** relaxes this restriction and can thus find a larger number of colors in CIELUV space that are spaced as far apart as possible. Some of these palettes have also been included in the base R function `palette.colors()` in **grDevices** (Zeileis *et al.* 2019) along with other qualitative palettes that provide more distinguishable colors than `qualitative_hcl()`. The palette collection packages **pals** (Wright 2019) and **paletteer** (Hvitfeldt 2020) also provide a wide range of prespecified palettes, including some qualitative schemes with many categories. Note that the palettes are quite diverse, though, and not all of them are equally suitable for coding qualitative information. The visualization functions in **colorspace** from Section 5 may be helpful in assessing their properties. **roloc** (Murrell 2018a,b) also provides color conversions, not between numeric color spaces, but rather from numeric color spaces to English color names.

In addition to the R version of **colorspace**, a Python 2/Python 3 (Van Rossum *et al.* 2011) re-implementation is available at `https://github.com/retostauffer/python-colorspace` which is currently in beta. In the paper we focus on the more mature R implementation replication materials for most examples are also available for Python.

Computational details

The results in this paper were obtained using R 4.0.3 (R Core Team 2020) with the packages **colorspace** 2.0-0 (Ihaka *et al.* 2020), **ggplot2** 3.3.2 (Wickham *et al.* 2020), **RColorBrewer** 1.1.2 (Neuwirth 2014), **rcartocolor** 2.0.0 (Nowosad 2019), **viridis** 0.5.1 (Garnier 2018), **scico** 1.2.0 (Pedersen and Crameri 2020). R itself and all packages used are available from the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/`.

Acknowledgments

The authors would like to thank the journal editors, the associate editor, and two anonymous reviewers for their constructive and helpful feedback that substantially improved the paper.

References

Aigner W (2010). "Perception and Visualization." URL http://www.ifs.tuwien.ac.at/~silvia/wien/vu-infovis/PDF-Files/02_perception-visualization_1up.pdf.

Brettel H, Viénot F, Mollon JD (1997). "Computerized Simulation of Color Appearance for Dichromats." *Journal of the Optical Society of America A*, **14**, 2647–2655. doi:10.1364/josaa.14.002647.

Brewer CA (1999). "Color Use Guidelines for Data Representation." In *Proceedings of the Section on Statistical Graphics, American Statistical Association*, pp. 55–60. Alexandria. URL http://www.personal.psu.edu/faculty/c/a/cab38/ColorSch/ASApaper.html.

CARTO (2019). "CARTOColors – Data-Driven Color Schemes." URL https://carto.com/carto-colors/.

Chang W, Cheng J, Allaire JJ, Xie Y, McPherson J (2020). *shiny: Web Application Framework for R*. R package version 1.5.0, URL https://CRAN.R-project.org/package=shiny.

Ciechanowski B (2019). "Color Spaces." URL https://ciechanow.ski/color-spaces/.

Coombes KR, Brock G (2020). *Polychrome: Qualitative Palettes with Many Colors*. R package version 1.2.5, URL https://CRAN.R-project.org/package=Polychrome.

Coombes KR, Brock G, Abrams ZB, Abruzzo LV (2019). "**Polychrome**: Creating and Assessing Qualitative Palettes with Many Colors." *Journal of Statistical Software, Code Snippets*, **90**(1), 1–23. doi:10.18637/jss.v090.c01.

Crameri F (2018). "Geodynamic Diagnostics, Scientific Visualisation and **StagLab** 3.0." *Geoscientific Model Development*, **11**(6), 2541–2562. doi:10.5194/gmd-11-2541-2018.

Gama J, Davis G (2018). ***colorscience**: Color Science Methods and Data*. R package version 1.0.5, URL https://CRAN.R-project.org/package=colorscience.

Garnier S (2018). ***viridis**: Default Color Maps from **matplotlib***. R package version 0.5.1, URL https://CRAN.R-project.org/package=viridis.

Harrower MA, Brewer CA (2003). "**ColorBrewer.org**: An Online Tool for Selecting Color Schemes for Maps." *The Cartographic Journal*, **40**(1), 27–37. doi:10.1179/000870403235002042. URL http://ColorBrewer.org/.

Hawkins E, McNeall D, Stephenson D, Williams J, Carlson D (2014). "The End of the Rainbow – An Open Letter to the Climate Science Community." URL http://www.climate-lab-book.ac.uk/2014/end-of-the-rainbow/.

Horvath M, Lipka C (2016). "sRGB Gamut within CIELCHuv Color Space Isosurface." Wikimedia Commons, URL https://commons.wikimedia.org/wiki/File:SRGB_gamut_within_CIELCHuv_color_space_isosurface.png.

Horvath M, Lipka C (2017). "sRGB Gamut within CIELCHuv Color Space Mesh." Wikimedia Commons, URL https://commons.wikimedia.org/wiki/File:SRGB_gamut_within_CIELCHuv_color_space_mesh.webm.

Hvitfeldt E (2020). *paletteer: Comprehensive Collection of Color Palettes*. R package version 1.2.0, URL https://CRAN.R-project.org/package=paletteer.

Ihaka R (2003). "Colour for Presentation Graphics." In K Hornik, F Leisch, A Zeileis (eds.), *Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria*. URL http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Proceedings/.

Ihaka R, Murrell P, Hornik K, Fisher JC, Stauffer R, Wilke CO, McWhite CD, Zeileis A (2020). **colorspace**: *A Toolbox for Manipulating and Assessing Colors and Palettes*. R package version 2.0-0, URL https://CRAN.R-project.org/package=colorspace.

Kaiser PK, Boynton RM (1996). *Human Color Vision*. 2nd edition. Optical Society of America, Washington.

Knoblauch K (2002). "Color Vision." In S Yantis, H Pashler (eds.), *Steven's Handbook of Experimental Psychology – Sensation and Perception*, volume 1, 3rd edition, pp. 41–75. John Wiley & Sons, New York.

Lumley T (2006). "Color Coding and Color Blindness in Statistical Graphics." *ASA Statistical Computing & Graphics Newsletter*, **17**(2), 4–7. URL http://stat-computing.org/newsletter/issues/scgn-17-2.pdf.

Lumley T (2013). **dichromat**: *Color Schemes for Dichromats*. R package version 2.0-0, URL https://CRAN.R-project.org/package=dichromat.

Machado GM, Oliveira MM, Fernandes LAF (2009). "A Physiologically-Based Model for Simulation of Color Vision Deficiency." *IEEE Transactions on Visualization and Computer Graphics*, **15**(6), 1291–1298. doi:10.1109/tvcg.2009.113. URL http://www.inf.ufrgs.br/~oliveira/pubs_files/CVD_Simulation/CVD_Simulation.html.

Murrell P (2018a). "Generating Colour Names: The **roloc** Package for R." URL https://stattech.wordpress.fos.auckland.ac.nz/2018/01/25/2018-01-generating-colour-names-the-roloc-package-for-r/.

Murrell P (2018b). **roloc**: *Convert Colour Specification to Colour Name*. R package version 0.1-1, URL https://CRAN.R-project.org/package=roloc.

Neuwirth E (2014). **RColorBrewer**: *ColorBrewer Palettes*. R package version 1.1-2, URL https://CRAN.R-project.org/package=RColorBrewer.

Nowosad J (2019). **rcartocolor**: *CARTOColors Palettes*. R package version 2.0.0, URL https://CRAN.R-project.org/package=rcartocolor.

Okabe M, Ito K (2008). "Color Universal Design (CUD): How to Make Figures and Presentations That Are Friendly to Colorblind People." URL http://jfly.iam.u-tokyo.ac.jp/color/.

Pedersen TL, Crameri F (2020). *scico: Colour Palettes Based on the Scientific Colour-Maps*. R package version 1.2.0, URL https://CRAN.R-project.org/package=scico.

Pedersen TL, Nicolae B, François R (2020). *farver: Vectorised Colour Conversion and Comparison*. R package version 2.0.3, URL https://CRAN.R-project.org/package=farver.

Poynton C (2009). "Frequently-Asked Questions about Color." URL http://www.poynton.com/ColorFAQ.html, accessed 2020-11-03.

Ram K, Wickham H (2018). ***wesanderson**: A Wes Anderson Palette Generator*. R package version 0.3.6, URL https://CRAN.R-project.org/package=wesanderson.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rost LC (2018). "What to Consider When Choosing Colors for Data Visualization." URL https://blog.datawrapper.de/colors/.

Smith N, Van der Walt S (2015). "A Better Default Colormap for **matplotlib**." In *SciPy 2015 – Scientific Computing with Python*. Austin. URL https://www.youtube.com/watch?v=xAoljeRJ3lU.

Stauffer R, Mayr GJ, Dabernig M, Zeileis A (2015). "Somewhere over the Rainbow: How to Make Effective Use of Colors in Meteorological Visualizations." *Bulletin of the American Meteorological Society*, **96**(2), 203–216. doi:10.1175/BAMS-D-13-00155.1.

Tufte ER (1990). *Envisioning Information*. Graphics Press, Cheshire.

Van Rossum G, *et al.* (2011). *Python Programming Language*. URL https://www.python.org/.

Viénot F, Brettel H, Mollon JD (1999). "Digital Video Colourmaps for Checking the Legibility of Displays by Dichromats." *Color Research and Application*, **24**(4), 243–252. doi:10.1002/(sici)1520-6378(199908)24:4<243::aid-col5>3.3.co;2-v.

Viénot F, Brettel H, Ott L, M'Barek AB, Mollon JD (1995). "What Do Colour-Blind People See?" *Nature*, **376**, 127–128. doi:10.1038/376127a0.

Ware C (1988). "Color Sequences for Univariate Maps: Theory, Experiments and Principles." *IEEE Computer Graphics and Applications*, **8**(5), 41–49. doi:10.1109/38.7760.

Ware C (2004). "Color." In *Information Visualization: Perception for Design*, chapter 4, pp. 103–149. Morgan Kaufmann Publishers.

Wickham H (2016). ***ggplot2**: Elegant Graphics for Data Analysis*. 2nd edition. Springer-Verlag. doi:10.1007/978-0-387-98141-3.

Wickham H, Chang W, Henry L, Pedersen TL, Takahashi K, Wilke CO, Woo K, Yutani H, Dunnington D (2020). *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. R package version 3.3.2, URL https://CRAN.R-project.org/package=ggplot2.

Wikipedia (2020a). "CIE 1931 Color Space — Wikipedia, The Free Encyclopedia." URL https://en.wikipedia.org/wiki/CIE_1931_color_space, accessed 2020-11-03.

Wikipedia (2020b). "CIELAB Color Space — Wikipedia, The Free Encyclopedia." URL https://en.wikipedia.org/wiki/CIELAB_color_space, accessed 2020-11-03.

Wikipedia (2020c). "CIELUV — Wikipedia, The Free Encyclopedia." URL https://en.wikipedia.org/wiki/CIELUV, accessed 2020-11-03.

Wikipedia (2020d). "Color Space — Wikipedia, The Free Encyclopedia." URL https://en.wikipedia.org/wiki/Color_space, accessed 2019-03-11.

Wikipedia (2020e). "HCL Color Space — Wikipedia, The Free Encyclopedia." URL https://en.wikipedia.org/wiki/HCL_color_space, accessed 2019-08-23.

Wikipedia (2020f). "HSL and HSV — Wikipedia, The Free Encyclopedia." URL https://en.wikipedia.org/wiki/HSL_and_HSV, accessed 2020-11-03.

Wikipedia (2020g). "RGB Color Space — Wikipedia, The Free Encyclopedia." URL https://en.wikipedia.org/wiki/RGB_color_space, accessed 2020-11-03.

Wikipedia (2020h). "sRGB — Wikipedia, The Free Encyclopedia." URL https://en.wikipedia.org/wiki/sRGB, accessed 2020-11-03.

Wikipedia (2020i). "Web Colors — Wikipedia, The Free Encyclopedia." URL https://en.wikipedia.org/wiki/Web_colors, accessed 2019-03-11.

Wilke CO (2019). *Fundamentals of Data Visualization*. O'Reilly Media. URL https://clauswilke.com/dataviz/color-basics.html.

Wilkinson L (2005). *The Grammar of Graphics*. 2nd edition. Springer-Verlag.

Wright K (2019). *pals: Color Palettes, Colormaps, and Tools to Evaluate Them*. R package version 1.6, URL https://CRAN.R-project.org/package=pals.

Zeileis A, Gaslam B, Murrell P, Pedersen TL (2018). "Benchmarking Color Space Conversions." Twitter discussion, URL https://twitter.com/AchimZeileis/status/1076228936810590208.

Zeileis A, Hornik K, Murrell P (2009). "Escaping RGBland: Selecting Colors for Statistical Graphics." *Computational Statistics & Data Analysis*, **53**, 3259–3270. doi:10.1016/j.csda.2008.11.033.

Zeileis A, Murrell P (2019). "HCL-Based Color Palettes in **grDevices**." The R Blog, URL https://developer.R-project.org/Blog/public/2019/04/01/hcl-based-color-palettes-in-grdevices/.

Zeileis A, Murrell P, Maechler M, Sarkar D (2019). "A New `palette()` for R." The R Blog, URL https://developer.R-project.org/Blog/public/2019/11/21/a-new-palette-for-r/.

Zhang S (2015). "Finding the Right Color Palettes for Data Visualizations." URL https://blog.graphiq.com/finding-the-right-color-palettes-for-data-visualizations-fcd4e707a283.

Affiliation:

Achim Zeileis
Universität Innsbruck
Department of Statistics
Faculty of Economics and Statistics
Universitätsstr. 15
6020 Innsbruck, Austria
E-mail: Achim.Zeileis@R-project.org
URL: https://eeecon.uibk.ac.at/~zeileis/

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
November 2020, Volume 96, Issue 1 *Submitted:* 2019-03-18
doi:10.18637/jss.v096.i01 *Accepted:* 2019-12-02

Article IX

Lang M.N., Lerch S., Mayr G.J., Simon T., Stauffer R., and Zeileis A. (2020). *Remember the Past: A Comparison of Time-Adaptive Training Schemes for Non-Homogeneous Regression*. Nonlinear Processes in Geophysics, 27, 23–34, doi:10.5194/npg-27-23-2020.

JCR ranking: **Category 2** in *Meteorology & Atmospheric Sciences*.

Contribution (CRT): *Conceptualization / data curation / software / formal analysis / validation / supervision / writing, original draft*.

Nonlin. Processes Geophys., 27, 23–34, 2020
https://doi.org/10.5194/npg-27-23-2020

Nonlinear Processes in Geophysics
Open Access
EGU

Remember the past: a comparison of time-adaptive training schemes for non-homogeneous regression

Moritz N. Lang[1,2], **Sebastian Lerch**[3], **Georg J. Mayr**[2], **Thorsten Simon**[1,2], **Reto Stauffer**[1,4], **and Achim Zeileis**[1]

[1]Department of Statistics, Universität Innsbruck, Innsbruck, Austria
[2]Department of Atmospheric and Cryospheric Sciences, Universität Innsbruck, Innsbruck, Austria
[3]Institute for Stochastics, Karlsruher Institut für Technologie, Karlsruhe, Germany
[4]Digital Science Center, Universität Innsbruck, Innsbruck, Austria

Correspondence: Moritz N. Lang (moritz.lang@uibk.ac.at)

Received: 27 September 2019 – Discussion started: 2 October 2019
Revised: 10 December 2019 – Accepted: 6 January 2020 – Published: 5 February 2020

Abstract. Non-homogeneous regression is a frequently used post-processing method for increasing the predictive skill of probabilistic ensemble weather forecasts. To adjust for seasonally varying error characteristics between ensemble forecasts and corresponding observations, different time-adaptive training schemes, including the classical sliding training window, have been developed for non-homogeneous regression. This study compares three such training approaches with the sliding-window approach for the application of post-processing near-surface air temperature forecasts across central Europe. The predictive performance is evaluated conditional on three different groups of stations located in plains, in mountain foreland, and within mountainous terrain, as well as on a specific change in the ensemble forecast system of the European Centre for Medium-Range Weather Forecasts (ECMWF) used as input for the post-processing.

The results show that time-adaptive training schemes using data over multiple years stabilize the temporal evolution of the coefficient estimates, yielding an increased predictive performance for all station types tested compared to the classical sliding-window approach based on the most recent days only. While this may not be surprising under fully stable model conditions, it is shown that "remembering the past" from multiple years of training data is typically also superior to the classical sliding-window approach when the ensemble prediction system is affected by certain model changes. Thus, reducing the variance of the non-homogeneous regression estimates due to increased training data appears to be more important than reducing its bias by adapting rapidly to the most current training data only.

1 Introduction

The need for accurate probabilistic weather forecasts is steadily increasing, because reliable information about the expected uncertainty is crucial for optimal risk assessment in agriculture and industry or for personal planning of outdoor activities. Therefore, most forecast centers nowadays issue probabilistic forecasts based on ensemble prediction systems (EPSs). To quantify the uncertainty of a specific forecast, an EPS provides a set of numerical weather predictions using slightly perturbed initial conditions and different model parameterizations (Palmer, 2002). However, due to various constraints and required simplifications in the EPS, these forecasts often show systematic biases and capture only parts of the expected uncertainty, especially when EPS forecasts are directly compared to point measurements (Gneiting and Katzfuss, 2014). In order to increase the predictive skill of the forecasts for specific locations, statistical post-processing is often applied to correct for these systematic errors in the forecasts' expectation and uncertainty.

One of the most frequently used parametric post-processing methods is "ensemble model output statistics" (EMOS) introduced by Gneiting et al. (2005). To emphasize that not only the errors in the mean but also the errors in the uncertainty are corrected, the method is often referred to as "non-homogeneous regression" (NR). In the statistical literature, this type of model is also known as distributional regression (Klein et al., 2014) since all parameters of a specific response distribution are optimized simultaneously conditional on respective sets of covariates.

As the error characteristics between the covariates, typically provided by the EPS, and the observations often show seasonal dependencies and might change inter-annually over time, different time-adaptive training schemes have been developed for NR models. Gneiting et al. (2005) proposed the so-called "sliding training window" approach where the training data set consists of EPS forecasts and observations of the most recent 30–60 d only. As soon as new data become available, the training data set and the statistical model are updated so that the estimated coefficients automatically evolve over time and adjust to changing error characteristics. This makes it very handy for operational use; however, little training data can sometimes yield unrealistic jumps in the estimated coefficients over time, especially if events which show a significantly different error characteristic enter the training data set. Therefore, to stabilize the temporal variability of the coefficient estimates, several approaches have been proposed in the literature. Scheuerer (2014) regularizes the estimation by only allowing the optimizer to slightly adjust the coefficient from day to day. In an alternative approach, Möller et al. (2018) extend the training data by using not only the days prior to estimation, but also the days centered around the same calendar day over all previous years available. This idea of using a rolling centered training data set over multiple years is similar to the concept of using annual cyclic smooth functions to capture seasonality as employed by Lang et al. (2019). These smooth functions are also known as regression splines (Wood, 2017), where the estimate of each point in the function only depends on data in its closer neighborhood; this allows for a smooth and stable evolution of the coefficients over the year.

Alternative time-adaptive models are based on historical analogs or non-parametric approaches. For approaches employing analogs (Junk et al., 2015; Barnes et al., 2019), training sets are selected to consist of past forecast cases with atmospheric conditions similar to those on the day of interest. Such methods may lead to models that are able to account for the flow dependency of EPS errors (Pantillon et al., 2018; Rodwell et al., 2018). However, the definition and computation of similarity measures are far from straightforward, and substantial methodological developments may be required to obtain suitably extensive training data sets for stable model estimation (Hamill et al., 2008; Lerch and Baran, 2017). For non-parametric approaches (Taillardat et al., 2016; Henzi et al., 2019) or semi-parametric approaches (Rasp and Lerch, 2018; Schlosser et al., 2019), time-adaptive choices of the training data are typically abandoned as well, as interactions between the day of the year and other covariates can capture the potential time adaptiveness. Therefore, analog-based and non-parametric approaches will not be pursued further in the context of this work.

In addition to the training scheme employed, an important data-specific aspect which has to be considered in post-processing is that the EPS may change over time (Hamill, 2018). This also motivates the recent study of Demaeyer and Vannitsem (2019), which introduces the promising concept of a post-processing method specifically dealing with model changes in a simplified physical setup. However, as stated by the authors, more research would be required to transfer their findings to real case scenarios. When using data of an operational EPS, changes in the underlying numerical model, e.g., an increased horizontal resolution, can typically lead to sudden transitions in the predictive performance of the EPS and hence affect the error characteristics of the data. If the training data set used to estimate the statistical post-processing model contains data of a previous EPS version which significantly differs from the current one, it can result in a loss of the predictive performance.

This paper presents a comparison of four widely used different time-adaptive training schemes proposed in the literature that employ alternative strategies to account for varying error characteristics in the data. To show a wide spectrum of possible approaches in a unified setup – rather than finding the universally best method – we consider typical basic applications of these training schemes and refrain from more elaborate tuning or combinations. A case study is shown for post-processed 2 m temperature forecasts for three different groups of stations across central Europe at the midlatitudes, namely, stations in the plain, in the foreland, and within mountainous terrain (Fig. 1). The study highlights the advantages and drawbacks of the different approaches in different topographical environments and investigates the impact of a change in the horizontal resolution of the EPS, which is expected to have a particularly pronounced effect on the predictive performance.

The structure of the paper is as follows: Sect. 2 explains the different methods and the comparison setup including the underlying data. In Sect. 3, the different time-adaptive training schemes are compared in terms of their coefficient paths and their predictive performance. Finally, a summary and conclusion are given in Sect. 4.

2 Methodology and comparison setup

The different training schemes for NR models proposed in the literature try to adapt to various kinds of error sources that can occur in post-processing, both in space and time. In order to provide a unifying view and to fix jargon, we first discuss these different error sources and then introduce the training schemes considered along with the comparison setup employed.

2.1 Sources of errors in post-processing

NR models aim to adjust for errors and biases in EPS forecasts but, of course, the NR models can be affected by errors and misspecifications themselves. Therefore, we try to carefully distinguish between the two different models involved with their associated errors, i.e., the numerical weather prediction model underlying the EPS vs. the statistical NR model employed for post-processing.

The skill of the EPS can be quantified in EPS forecast biases and variances, which (i) typically vary for different locations conditional on the surrounding terrain, (ii) often show cyclic seasonal patterns, and (iii) can experience non-seasonal temporal changes, e.g., due to changes in the EPS itself.

In addition to the error sources in the employed EPS, the performance of the statistical post-processing itself will typically also (iv) differ at different measurement sites, (v) strongly depend on the amount of training data used, and (vi) whether it is affected by effects that are not accounted for in the NR specification.

Clearly, larger training samples (v) will lead to more reliable predictions when the NR specification (vi) – in terms of response distribution, covariates and corresponding effects, link functions, estimation method, etc. – appropriately captures the error characteristics in the relationship between EPS forecasts and actual observations. However, when these error characteristics differ in space (i and iv) and/or in time (ii and iii), it is not obvious what the best strategy for training the NR is. Extending the training data (v) in space or time will reduce the variance of the NR estimation but might also introduce bias if the NR specification (vi) is not adapted. Thus, this is a classical bias-variance trade-off problem, and we investigate which strategies for dealing with this are most useful in a typical temperature forecasting situation.

To fix jargon, we employ the terms "model" and "bias" without further qualifiers when referring to the NR model in post-processing, whereas when referring to the numerical weather prediction model we employ "EPS model" and "EPS bias". Moreover, we refer to a statistical model whose estimates have small bias and variance as stable.

2.2 Non-homogeneous regression with time-adaptive training schemes

Non-homogeneous regression as originally introduced by Gneiting et al. (2005) is a special case of distributional regression, where a response variable y is assumed to follow a specific probability distribution $\mathcal{D}$ with distribution parameters $\theta_k, k = 1, \ldots, K$:

$$y \sim \mathcal{D}(\theta_1, \ldots, \theta_K) = \mathcal{D}(h_1(\eta_1), \ldots, h_K(\eta_K)), \tag{1}$$

where each parameter of the distribution is linked to an additive predictor η_k via a link function h_k to ensure its appropriate co-domain. In the case of post-processing air temperatures, the normal distribution is typically employed (Gneiting and Katzfuss, 2014), and Eq. (1) can be rewritten as

$$y \sim \mathcal{N}(\mu, \sigma). \tag{2}$$

In the classical NR (Gneiting et al., 2005), the two distribution parameters location μ and scale σ are expressed by the ensemble mean m and ensemble variance or standard deviation s, respectively:

$$\mu = \eta_\mu = \beta_0 + \beta_1 \cdot m, \tag{3}$$

$$\log(\sigma) = \eta_\sigma = \gamma_0 + \gamma_1 \cdot s, \tag{4}$$

with $\beta_\bullet$ and $\gamma_\bullet$ being the corresponding intercept and slope coefficients. Here, we use the logarithm link to ensure positivity of the scale parameter σ; however, a quadratic link with additional parameter constraints for the coefficients as used by Gneiting et al. (2005) would also be feasible. In this study, we regard the statistical model specifications according to Eqs. (2)–(4), but all concepts of time-adaptive training schemes could easily be transferred to other response distributions $\mathcal{D}$, to alternative link functions $h(\cdot)$, or to more complex additive predictors η with additional covariates.

The regression coefficients $\beta_\bullet$ and $\gamma_\bullet$ are estimated by minimizing a loss function over a training data set containing historical pairs of observations and EPS forecasts. In this study, we employ maximum likelihood estimation, which performs very similarly to minimizing the continuous ranked probability score (CRPS, Gneiting and Raftery, 2007) as used by Gneiting et al. (2005) when the response distribution is well specified (Gebetsberger et al., 2018). For a single observation y, the log-likelihood L of the normal distribution is given by

$$L(\mu, \sigma | y) = \log\left\{\frac{1}{\sigma}\,\phi\left(\frac{y-\mu}{\sigma}\right)\right\}, \tag{5}$$

where $\phi(\cdot)$ is the probability density function of the normal distribution. The coefficients $\beta_\bullet$ and $\gamma_\bullet$, specified in Eqs. (3) and (4), are derived by minimizing the sum of negative log-likelihood contributions L over the training data. The larger the training data, the more stable the estimation in case the statistical model is well specified; however, if the covariate's skill varies either seasonally or non-seasonally over time, this leads to the bias-variance trade-off between preferable large training data sets for stable estimation and the benefit of shorter training periods which allow one to adjust more rapidly to changes in the data or, to be precise, in the error characteristics of the data (see Sect. 2.1). In the following, four approaches are discussed on how to gain informative time-adaptive training data sets while ensuring a stable estimation.

2.2.1 Sliding-window

The *sliding-window* approach originally introduced by Gneiting et al. (2005) uses the most recent days prior to

the day of interest as training data for estimation. For post-processing 2 m temperature forecasts, Gneiting et al. (2005) found the best predictive performance for training periods between 30 and 45 d with substantial gains in increasing the training period beyond 30 d and slow but steady performance losses for training lengths beyond 45 d. According to Gneiting et al. (2005), the latter is presumably a result of seasonally varying EPS forecast biases.

In this study, we use a period of 40 d for the *sliding-window* approach, which is a frequently used compromise (e.g., Baran and Möller, 2017; Gneiting et al., 2005; Wilson et al., 2007). However, as discussed in Gneiting et al. (2005), different training periods might perform better for distinct weather variables, locations, forecast steps, or model specifications. Common choices in the literature include training lengths between 15 and 100 d, for example, depending on whether the estimation of regression coefficients is performed station-specifically or jointly for multiple locations at once.

2.2.2 Regularized sliding-window

A regularized adaption of the classical *sliding-window* approach was introduced by Scheuerer (2014) in order to stabilize the estimation based on early stopping in statistical learning. The motivation is that gradient-based optimizers adjust the starting values by iteratively taking steps in the direction of the steepest descent of a distinct loss function until some convergence condition is fulfilled. These steps are largest in the first iteration and get smaller towards the optimum. Thus, the most important adjustments are made during the first steps, while further adjustments often improve the fit to unimportant or even random features in the data, which can lead to wiggly coefficient paths over time and ultimately to an overfitting (Scheuerer, 2014).

Therefore, Scheuerer (2014) proposes to use the coefficients of the previous day as starting values and to stop the optimizer after a single iteration to stabilize the evolution of the coefficient estimates. A drawback of his approach is that it implies that the estimation never converges and, in the case of poor starting values or strong truly observed temporal changes in the data, the obtained coefficients might be incorrect (Scheuerer, 2014). For post-processing precipitation amounts employing a left-censored generalized extreme value distribution, Scheuerer (2014) obtained better results with regularized coefficients than without regularization.

For the *regularized sliding-window* approach used in this study, we employ the quasi-Newton Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm as in Scheuerer (2014) and stop the optimizer after one single iteration. For the first time, we let the BFGS algorithm perform 10 iterations and use $(\beta_0, \beta_1)^\top = (0, 1)^\top$ as starting values in the location parameter μ and $(\gamma_0, \gamma_1)^\top = (0.1, 1)^\top$ as starting values in the scale parameter σ. According to Scheuerer (2014) a single iteration might not always provide the optimum degree of regularization; however, for the presented comparison study a single iteration yields a regularized setup which is on the opposite side of the possible model spectrum compared to the classical *sliding-window* approach which runs until convergence. In comparison to Scheuerer (2014), we perform maximum likelihood estimation instead of CRPS minimization.

2.2.3 Sliding-window plus

As already pointed out by Gneiting et al. (2005), training data from previous years could additionally be included in the *sliding-window* approach to address seasonal effects. This should reduce the variance in the estimation of the regression coefficients, which stabilizes the evolution of the coefficients similarly to the *regularized sliding-window* approach.

This idea has recently been pursued by Vogel et al. (2018) for the construction of climatological reference forecasts and by Möller et al. (2018) for a post-processing approach based on D-vine copulas in which many more coefficients than in classical NR need to be estimated, making a more extensive training data set necessary. Their so-called "refined training data set" consists of the most 45 recent days prior to the day of interest plus 91 d centered around the same calendar day over all previous years available. Including multiple years yields more stable estimates, while, on the other hand, there is the trade-off of losing the ability to quickly adjust to non-seasonal temporal changes in the EPS forecast biases. The approach of Möller et al. (2018) can be seen as time-adaptive version of the seasonal training proposed by Hemri et al. (2016), who consider training data sets comprised of days from all previous years within the same season (winter/summer).

In this study, to be comparable to the *sliding-window* approach, we use the most recent 40 d prior to estimation and a respective 81 d interval centered around the day of interest over the previous years available in the training data.

2.2.4 Smooth model

If we reformulate the *sliding-window plus* approach, it is very similar to fitting an annual cyclic smooth function where the points of the function only depend on data points in the closer neighborhood, specified by the sliding-window length.

Cyclic smooth functions belong to the broader model class of generalized additive models (GAMs, Hastie and Tibshirani, 1986), which allow one to include potentially nonlinear effects in the linear predictors η. Smooth functions are also referred to as regression splines and are directly linked to the model parameters as additive terms in η. Introductory material for cyclic smooth functions conditional on the day of the year can be found in Lang et al. (2019), and a comprehensive summary of GAMs is given in Wood (2017).

To account for seasonal variations we only need to fit one single model, here called the *smooth model*, over a training

data set with several years of data. The effects included allow the coefficients to smoothly evolve over the year, which leads to the following adaptations in Eq. (3) and (4) for the location μ and scale σ, respectively:

$$\mu = \eta_\mu = \beta_0 + f_0(\text{doy}) + (\beta_1 + f_1(\text{doy})) \cdot m, \quad (6)$$

$$\log(\sigma) = \eta_\sigma = \underbrace{\gamma_0 + g_0(\text{doy})}_{\substack{\text{seasonally varying} \\ \text{intercept}}} + \underbrace{(\gamma_1 + g_1(\text{doy}))}_{\substack{\text{seasonally varying} \\ \text{slope}}} \cdot s, \quad (7)$$

with m and s being the ensemble mean and ensemble standard deviation, respectively; $\beta_\bullet$ and $\gamma_\bullet$ are regression coefficients, and $f_\bullet(\text{doy})$ and $g_\bullet(\text{doy})$ employ cyclic regression splines conditional on the day of the year (Wood, 2017). The regression coefficients β_0 and γ_0, as well as β_1 and γ_1, are unconditional on the day of the year and can be interpreted as global intercept or slope coefficients, respectively.

2.3 Comparison setup

2.3.1 NR training schemes

The NR training schemes presented in Sect. 2.2 deal with the potential temporal error sources from Sect. 2.1 in different ways (see Table 1 for an overview). The classic *sliding-window* employs the basic NR model equations from Eqs. (3) to (4) and avoids potential biases in the NR model estimation by using only very recent data from the same year and season. Compared to this, the *regularized sliding-window* and *sliding-window plus* approaches both try to stabilize the coefficient estimates by reducing the variance – either through regularized estimation (vi) or by considering multiple years (v). The *smooth model* differs from all of these by modifying both the model (vi) and data (v) specification, using the extended model specification from Eqs. (6) to (7) fitted by penalized estimation to a large data set comprising several years and all seasons.

Potential spatial differences (i) and (iv) are handled for all training schemes in the same way: the NR models are estimated separately for each station and subsequently evaluated in groups of terrain types (plain, foreland, alpine). The underlying EPS data – described subsequently – are the same for all NR training schemes and are thus affected by the same seasonal (ii) and non-seasonal changes (iii).

2.3.2 Data sets

For validation of the training schemes, we consider 2 m temperature ensemble forecasts and corresponding observations at 15 measurement sites located across Austria, Germany, and Switzerland. The sites are chosen to investigate the impact of potential error sources in space (i) and (iv), e.g., through varying discrepancies between the real and EPS topography. The data comprises three groups of five stations located either in plains, in mountain foreland, or within mountainous terrain (see Fig. 1). The estimated statistical models

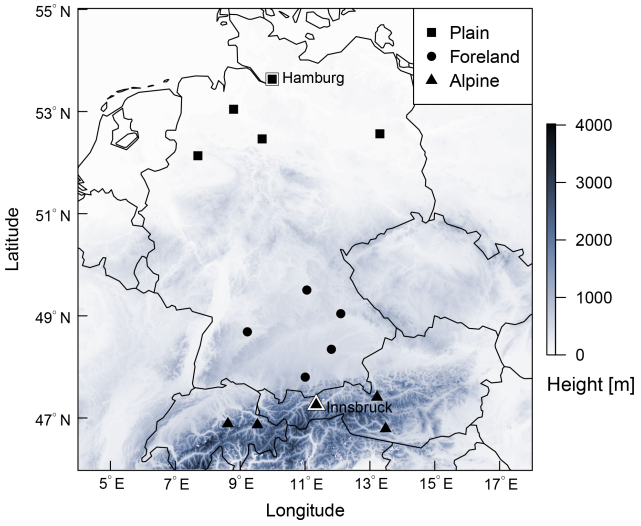

Figure 1. Overview of the study area with selected stations classified as plain, foreland, and alpine station sites. The two highlighted and labeled stations, Hamburg and Innsbruck, are discussed in detail in Sect. 3.1. Elevation data are obtained from the SRTM-30 m digital elevation model (NASA JPL, 2013).

for stations Hamburg and Innsbruck, highlighted by symbols with white borders, are discussed in more detail in Sect. 3.1.

As covariates for Eqs. (3)–(7), we employ the ensemble mean m and the ensemble standard deviation s of bilinearly interpolated 2 m temperature forecasts issued by the global 50-member EPS of the European Centre for Medium-Range Weather Forecasts (ECMWF). We assess forecast steps from +12 to +72 h ahead at a 12-hourly temporal resolution for the EPS run initialized at 00:00 UTC and use data from 8 March 2010 to 7 March 2019.

This period has been selected in order to investigate the impact of non-seasonal long-term changes in the EPS model (iii) that is not reflected in the NR model specifications; i.e., the horizontal resolution of the ECMWF EPS changed from the previous version (cycle 36r1; 26 January 2010) to the new version on 8 March 2016 (cycle 41r2). This specific model change was chosen among various others as it modifies the height of the terrain and, thus, likely introduces an EPS bias for temperature forecasts directly affecting the coefficient estimates; other changes such as modified model parameterizations or improvements in the analysis scheme are expected to have a minor impact on the post-processing of 2 m temperatures. It is of specific interest how the *sliding-window plus* and the *smooth model* are affected if the training period comprises data from both the "old EPS version" before the change in the horizontal resolution as well as the "new EPS version". Thus, we construct three data sets with different validation periods that are either (a) not affected by this EPS model change at all, (b) start immediately after the model change, or (c) have some time lag after change.

Table 1. Overview of time-adaptive training schemes, distinguished by model specification/estimation and training data selection corresponding to errors sources (vi) and (v), respectively. The basic model specification refers to Eqs. (3)–(4), in contrast to the extended Eqs. (6)–(7).

	Model		Data	
Name	Specification	Estimation	Years	Seasons
Sliding-window	Basic	Maximum likelihood	Current	Current
Regularized sliding-window	Basic	Early stopping	Current	Current
Sliding-window plus	Basic	Maximum likelihood	Multiple	Current
Smooth model	Extended	Penalized	Multiple	All

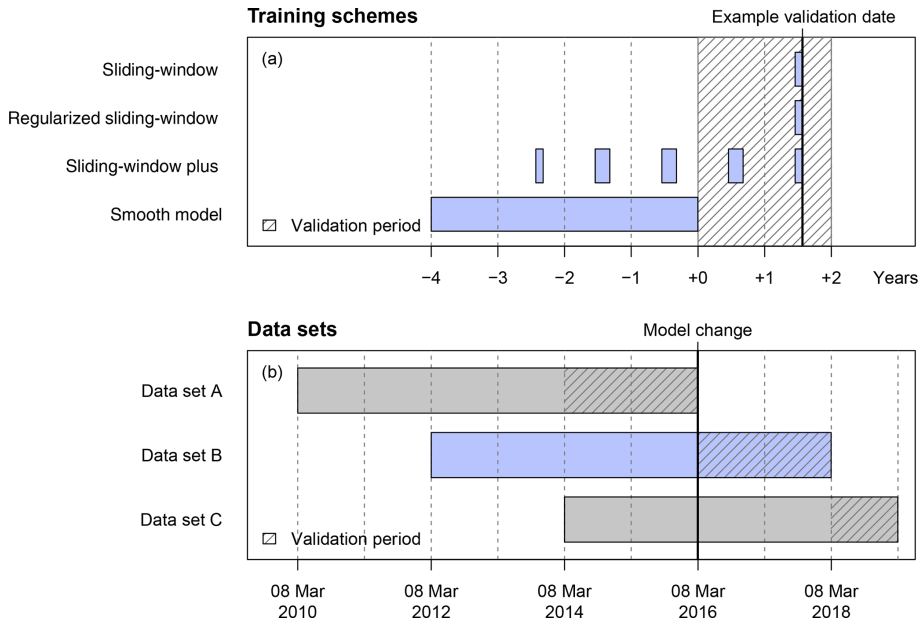

Figure 2. **(a)** Illustrative example of how the training data sets are composed for the four different time-adaptive training schemes. **(b)** Schematic overview of the training and validation data sets employed in this study with regard to the change in the horizontal resolution of the ECMWF EPS on 8 March 2016 (cycle 41r2). For training, up to 4 years of data are used in all data sets; for validation, 2 years of data are used for data sets A and B, and 1 year for data set C.

To understand how this affects the different training schemes, we first illustrate in Fig. 2a how training and validation periods are selected for each scheme. For the three sliding-window approaches, the NR models are re-estimated every day as the validation date rolls through the validation period (hatched area). In contrast, the *smooth model* is estimated only once for the entire validation period based on a fixed training data period of 4 years prior to the validation period. For a fair comparison, the training data for the *sliding-window plus* model are also restricted to 4 years prior to each validation date.

Now Fig. 2b illustrates how the three data sets A, B, and C are selected in relation to the EPS change on 8 March 2016.

- Data set A. All models are trained and evaluated without being affected by the EPS change.
- Data set B. All models start with a training period entirely before the EPS change but a validation period entirely after the change. However, for the *sliding-window* and *regularized sliding-window* approaches, the training period quickly rolls across the change point, and after 40 d they are not affected by it anymore. For *sliding-window plus* the training data also roll into the new EPS version but still partially use data from the old EPS version. Finally, as the *smooth model* is only estimated once, it cannot adapt at all to the new EPS version.
- Data set C. Effects from A and B are mixed so that the *smooth model* and the *sliding-window plus* model use data from both the old and new EPS versions, while the classical *sliding-window* and *regularized sliding-window* models already use only data from the new EPS version.

The validation period is 2 years for A and B and 1 year for C. A total number of 731/730/365 NR models has to be estimated for the three sliding-window approaches, while only

1/1/1 *smooth model* is required for data sets A/B/C per station and forecast step. The computation time for the various sliding-window approaches is in the order of seconds, whereas the estimation of the *smooth model*, including full Markov chain Monte Carlo (MCMC) sampling, is in the order of minutes on a standard computer.

3 Results

This section assesses the performance of the different time-adaptive training schemes. First, the temporal evolutions of the estimated coefficients are shown for two stations representative of one measurement site in the plains and one in mountainous terrain. Afterwards, the predictive performance of the training schemes is evaluated in terms of the CRPS conditional on the three data sets with and without the change in the horizontal resolution of the EPS (Fig. 2) and grouped for stations classified as topographically plain, mountain foreland, and alpine sites (Fig. 1).

3.1 Coefficient paths

Figure 3 shows the estimated coefficients for Innsbruck at forecast step +36 h conditional on the day of the year. The coefficient paths are plotted for the different time-adaptive training schemes for 2 years included in the validation period of data set A. The pronounced seasonal evolution of the coefficients for all training schemes shows that the EPS' forecast bias and skill varies seasonally, which makes a time-adaptive training scheme mandatory to capture these characteristics in the post-processing. During summer, a slope coefficient β_1 close to 1 in the location parameter μ and a high slope coefficient γ_1 in the scale parameter σ indicate a better performance of the EPS compared to the cold season.

In comparison to the other time-adaptive training schemes, the classical *sliding-window* approach (Fig. 3a, d, g, j) shows very strong outliers and an unstable temporal evolution for all coefficients with distinct differences during the 2 subsequent validation years; this is more pronounced for the scale parameter σ where the estimates seem to be more volatile than for the location parameter μ. All strategies extending the classical *sliding-window* approach smooth the temporal evolution of the coefficients to a certain extent while maintaining the overall seasonal cyclic pattern. For the *regularized sliding-window* approach (Fig. 3b, e, h, k), the stabilization strongly differs for the individual coefficients, and some of the estimated coefficients seem to need rather long to adapt during the transition periods; the latter could indicate that a single iteration step might not be sufficient in this study. The coefficient paths for the *sliding-window plus* approach (Fig. 3c, f, i, l) and for the *smooth model* (Fig. 3a–l; solid line) look very similar with minor distortions during the cold season. Due to the definition of the *smooth model*, its coefficient paths show the most stable evolution but with the lowest ability to react to abrupt changes in the error characteristics.

For Hamburg (Fig. 4) by contrast to Innsbruck, the information content of the mean EPS temperature forecast is quite high throughout the year. This yields a lower bias correction and an almost one-to-one mapping of the ensemble mean to the location parameter μ indicated by a coefficient β_1 close to 1. Despite the different post-processing characteristics, the temporal evolution of the coefficient paths is similar to the one for Innsbruck, which confirms our previous findings: for the extended sliding-window approaches the coefficients have indeed very little seasonal variability, while for the classical *sliding-window* approach the coefficients show unrealistically strong fluctuations over time without a clear seasonal pattern (Fig. 4a, d, g, j). As for Innsbruck, the *regularized sliding-window* approach has a rather unrealistic stepwise evolution for some coefficients (Fig. 4b, e, h, k). The coefficient paths for the *sliding-window plus* approach (Fig. 4c, f, i, l) and the *smooth model* (Fig. 4; solid line) look comparable. These results support the bias-variance trade-off where regularizing or smoothing stabilizes the coefficient paths while losing the ability to rapidly react to temporal changes in the data.

3.2 Predictive performance

After the illustrative evaluation of the coefficients' temporal evolution for the different time-adaptive training schemes, Fig. 5 shows aggregated CRPS skill scores for groups of five respective stations classified as topographically plain, mountain foreland, and alpine sites (Fig. 1) regarding data sets A, B, and C (Fig. 2). In all panels the *regularized sliding-window* approach, the *sliding-window plus* approach, and the *smooth model* are compared to the classical *sliding-window* approach as a reference.

- For data set A, the *regularized sliding-window* approach shows only little improvements for the plain and foreland and an overall performance loss for alpine stations. By contrast, the *sliding-window plus* and *smooth model* approaches show distinct improvements over the classical *sliding-window* approach, with the largest values for alpine sites.

- For data set B at stations in the plains and foreland, the mean predictive skill behaves similarly to data set A, except that the *smooth model* shows a slightly larger variance. For alpine stations, the *regularized sliding-window* approach performs slightly worse than in data set A, while the two approaches using training data over multiple years no longer outperform the reference.

- For data set C at stations in the plains and foreland, the predictive skill is again similar to data set A with slight performance losses. For alpine stations, the *regularized sliding-window* approach shows even less skill than in

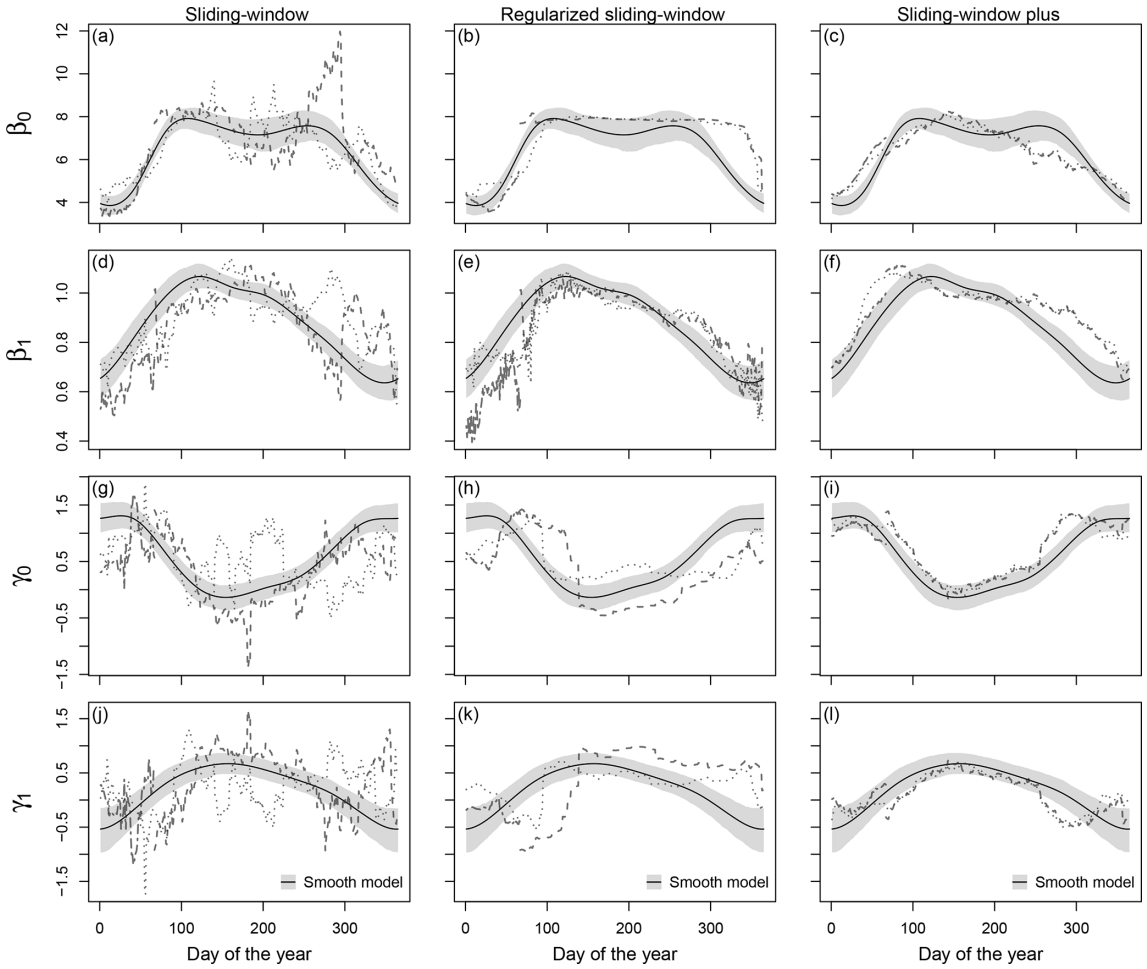

Figure 3. Temporal evolution of regression coefficients for the validation period in data set A for Innsbruck at forecast step +36 h (valid at 12:00 UTC). The coefficient paths are shown for the coefficients β_0 **(a–c)** and β_1 **(d–f)** in the location parameter μ and for the coefficients γ_0 **(g–i)** and γ_1 **(j–l)** in the scale parameter σ based on the *sliding-window*, *regularized sliding-window*, and *sliding-window plus* approaches (dashed, from left to right) compared to the *smooth model* approach (solid line). The coefficient paths are plotted for the consecutive calendar years 2014, 2015, and 2016 as dashed, dotted, and two-dashed lines, respectively. The grey shading represents the 95 % credible intervals of the coefficients in the *smooth model* based on MCMC sampling.

data set B, while the two other approaches again outperform the *sliding-window* approach and are on a similar level to that in data set A.

The validation of the different time-adaptive training schemes shows that the *sliding-window plus* approach and the *smooth model* perform overall similarly and are clearly superior for all station types compared to the classical *sliding-window* approach. However, the *smooth model* shows the highest variance in the predictive performance in the case of a change in the EPS, especially in mountainous terrain (data sets B and C); this is likely due to its reduced ability to adapt to temporal changes in the data. Furthermore, the validation shows that the *regularized sliding-window* approach seems to have difficulties in mountainous terrain and yields only minor improvements for plain and foreland sites.

4 Summary and conclusion

Non-homogeneous regression (NR) is a widely used method to statistically post-process ensemble weather forecasts. In its original version it was used for temperature forecasts employing a Gaussian response distribution, but over the last decade various statistical model extensions have been proposed for other quantities employing different response distributions or to enhance its predictive performance. When estimating NR models there is always a trade-off between large enough training data sets to get stable estimates and still allowing the statistical model to adjust to temporal changes in the statistical error characteristics of the data. Therefore, different training schemes with specific advantages and drawbacks have been developed as presented in this paper. To show a wide spectrum of possible approaches in a unified

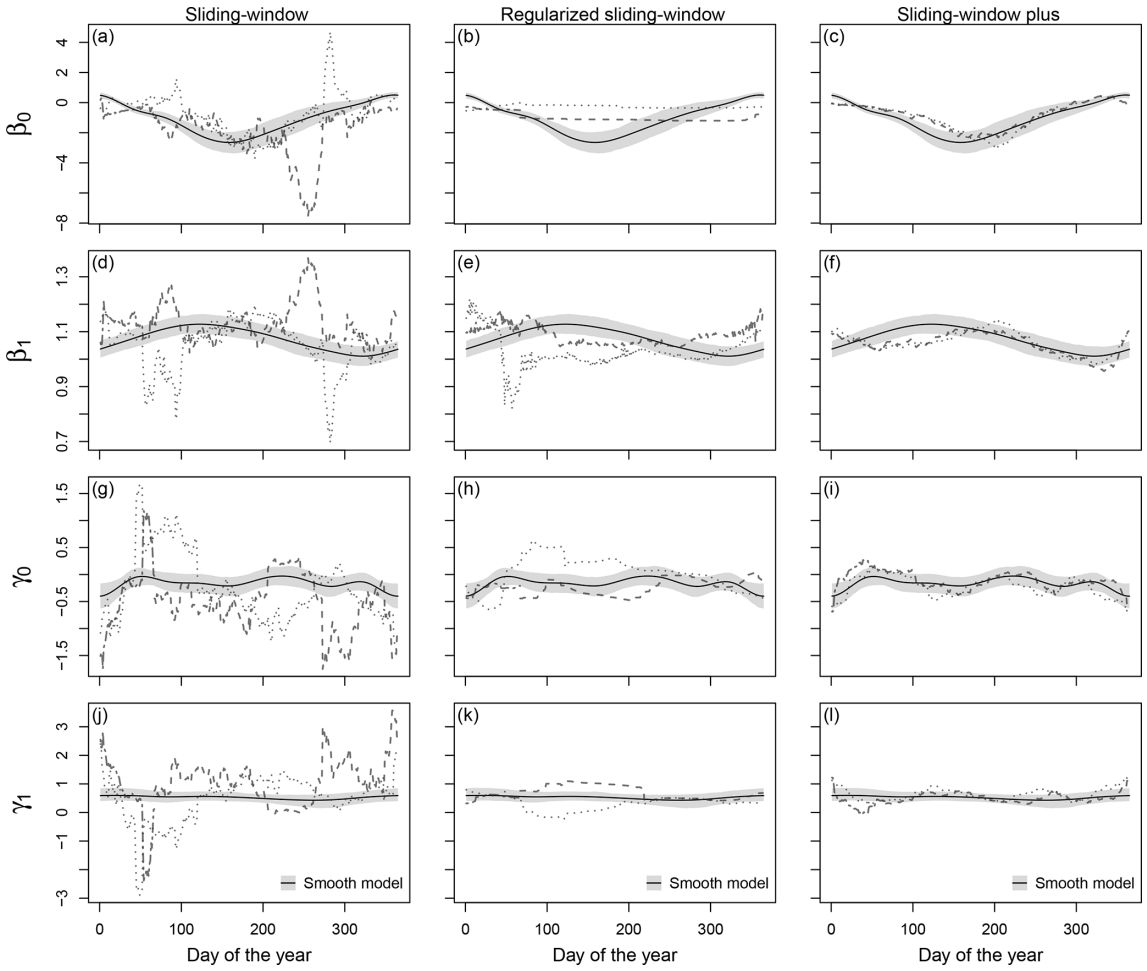

Figure 4. As Fig. 3 but for Hamburg at forecast step +36 h (valid at 12:00 UTC).

setup, we consider typical basic applications of the training schemes and refrain from more elaborate tuning or combinations.

The classical *sliding-window* approach has the advantage that no extensive training data set is required, which allows the statistical model to adjust itself rapidly to changing forecast biases, for example, in the case of changes in the EPS. On the other hand, statistical models trained on a small training data set have typically large variance in the estimation of the regression coefficients, which can yield unstable wiggly coefficient paths. Additional regularization allows one to stabilize the evolution of the regression coefficients without losing the simplicity of the classical *sliding-window* approach. However, inappropriate settings of the optimizer, e.g., unrealistic starting values or insufficient update steps, can quickly lead to incorrect coefficients. The alternative *sliding-window plus* strategy foregoes regularization but stabilizes the coefficients by using an extended training data set which includes data from the same season over several years. Compared to the classical approach the method requires historical data and partially loses its ability to rapidly adjust to changes in the error characteristics. The last approach presented in this paper can be seen as a generalization of the *sliding-window plus* approach. Rather than using a training data set centered around the date of interest, the *smooth model* makes use of all historical data in combination with cyclic regression splines, which allows the coefficients to smoothly evolve over the year.

The differences between the methods presented can be seen in the coefficient paths shown in Figs. 3 and 4. The coefficients of the classical *sliding-window* approach show strong fluctuations and pronounced peaks throughout the year. Regularization allows one to stabilize the evolution; however, strong step-wise changes in the coefficient paths still occur. The two methods using data from multiple years perform comparably similarly and show stable coefficient paths over the year. Figure 5 confirms that more stable estimates have a positive impact on the predictive performance. The *sliding-window plus* approach and the *smooth model* show an overall improvement of about 3 %–5 % (in median) over the classical *sliding-window* approach, while the *regularized sliding-window* only partially outperforms the *sliding-window* training scheme. Even in the case of the model change chosen to

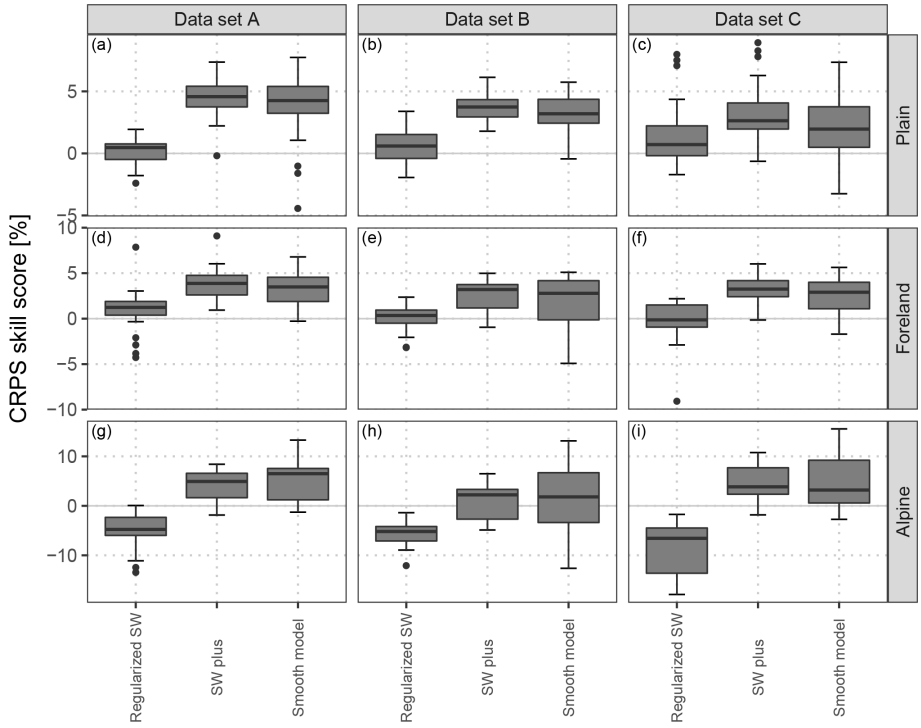

Figure 5. CRPS skill scores clustered into groups of stations located in the plain, in the mountain foreland near the Alps, and within mountainous terrain and for the out-of-sample validation periods according to the different data sets: data set A without the change in the horizontal resolution of the EPS, data set B with the EPS change in between the training and validation data sets, and data set C with the EPS change within training data (Fig. 2). Compared are the different time-adaptive training schemes specified in Sect. 2.2 with the classical *sliding-window* approach as a reference; note that "sliding-window" is abbreviated as SW in the figure. Each box-and-whisker contains aggregated skill scores over the forecast steps from +12 to +72 h at a 12-hourly temporal resolution and over five respective weather stations (Fig. 1). Skill scores are in percent and positive values indicate improvements over the reference.

demonstrate the effect of non-seasonal long-term changes on the coefficient estimates, the training schemes using multiple years of data are still superior to the ones using the most recent days only, even if they technically allow adjustment to the EPS change more rapidly.

To conclude, all four training schemes shown in this paper have their advantages in particular applications. If only short periods of training data are available (< 1 year), the classical *sliding-window* approach may already provide sufficiently good estimates. However, as soon as one has access to longer historical data sets, the approaches using data from multiple years become superior due to a more stable coefficient evolution over time, which yields an overall improved performance. This even holds in the case of the EPS change considered in this study, but may be different for other changes or EPSs. While the *sliding-window plus* approach is a natural extension of the classical *sliding-window* approach and, therefore, can be estimated by the same software, the *smooth model* approach can be seen as a generalization, and only a single model has to be estimated for all seasons using all available data. The *smooth model* yields, by definition, the smoothest and most stable coefficient paths but with the lowest ability to adjust itself to a new error characteristic.

Code availability. All computations are performed in R 3.6.1 (R Core Team, 2019) https://www.R-project.org/ (last access: 10 December 2019). The statistical models using a sliding-window approach are based on R package **crch** (Messner et al., 2016) (https://doi.org/10.32614/RJ-2016-012) employing a frequentist maximum likelihood approach. The statistical models using a time-adaptive training scheme by fitting cyclic smooth functions are fitted with R package **bamlss** (Umlauf et al., 2018) (https://doi.org/10.1080/10618600.2017.1407325). The package provides a flexible toolbox for distribution regression models in a Bayesian framework; introductory material can be found at http://BayesR.R-Forge.R-project.org/ (last access: 10 December 2019). The computation of the CRPS is based on R package **scoringRules** (Jordan et al., 2019) (https://doi.org/10.18637/jss.v090.i12).

Supplement. The supplement related to this article is available online at: https://doi.org/10.5194/npg-27-23-2020-supplement.

Author contributions. This study is based on the PhD work of MNL under supervision of GJM and AZ. The majority of the work for this study was performed by MNL with the support of RS. All the

authors worked closely together in discussing the results and commenting on the manuscript.

Competing interests. Sebastian Lerch is one of the editors of the special issue on "Advances in post-processing and blending of deterministic and ensemble forecasts". The remaining authors declare that they have no conflict of interest.

Special issue statement. This article is part of the special issue "Advances in post-processing and blending of deterministic and ensemble forecasts". It is not associated with a conference.

Acknowledgements. We thank the Zentralanstalt für Meteorologie und Geodynamik (ZAMG) for providing access to the data.

Financial support. This project was partly funded by the Austrian Research Promotion Agency (FFG, grant no. 858537) and by the Austrian Science Fund (FWF, grant no. P31836). Sebastian Lerch gratefully acknowledges support by the Deutsche Forschungsgemeinschaft (DFG) through SFB/TRR 165 "Waves to Weather".

Review statement. This paper was edited by Maxime Taillardat and reviewed by two anonymous referees.

References

Baran, S. and Möller, A.: Bivariate Ensemble Model Output Statistics Approach for Joint Forecasting of Wind Speed and Temperature, Meteorol. Atmos. Phys., 129, 99–112, https://doi.org/10.1007/s00703-016-0467-8, 2017.

Barnes, C., Brierley, C. M., and Chandler, R. E.: New approaches to postprocessing of multi-model ensemble forecasts, Q. J. Roy. Meteor. Soc., 145, 3479–3498, https://doi.org/10.1002/qj.3632, 2019.

Demaeyer, J. and Vannitsem, S.: Correcting for Model Changes in Statistical Post-Processing – An approach based on Response Theory, Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2019-57, in review, 2019.

Gebetsberger, M., Messner, J. W., Mayr, G. J., and Zeileis, A.: Estimation Methods for Nonhomogeneous Regression Models: Minimum Continuous Ranked Probability Score versus Maximum Likelihood, Mon. Weather Rev., 146, 4323–4338, https://doi.org/10.1175/MWR-D-17-0364.1, 2018.

Gneiting, T. and Katzfuss, M.: Probabilistic Forecasting, Annu. Rev. Stat. Appl., 1, 125–151, https://doi.org/10.1146/annurev-statistics-062713-085831, 2014.

Gneiting, T. and Raftery, A. E.: Strictly Proper Scoring Rules, Prediction, and Estimation, J. Am. Stat. Assoc., 102, 359–378, https://doi.org/10.1198/016214506000001437, 2007.

Gneiting, T., Raftery, A. E., Westveld III, A. H., and Goldman, T.: Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation, Mon. Weather Rev., 133, 1098–1118, https://doi.org/10.1175/MWR2904.1, 2005.

Hamill, T. M.: Practical Aspects of Statistical Postprocessing, in: Statistical Postprocessing of Ensemble Forecasts, edited by: Vannitsem, S., Wilks, D. S., and Messner, J. W., 187–217, Elsevier, https://doi.org/10.1016/C2016-0-03244-8, 2018.

Hamill, T. M., Hagedorn, R., and Whitaker, J. S.: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation, Mon. Weather Rev., 136, 2620–2632, https://doi.org/10.1175/2007MWR2411.1, 2008.

Hastie, T. and Tibshirani, R.: Generalized Additive Models, Stat. Sci., 1, 297–310, 1986.

Hemri, S., Haiden, T., and Pappenberger, F.: Discrete Postprocessing of Total Cloud Cover Ensemble Forecasts, Mon. Weather Rev., 144, 2565–2577, https://doi.org/10.1175/mwr-d-15-0426.1, 2016.

Henzi, A., Ziegel, J. F., and Gneiting, T.: Isotonic Distributional Regression, arXiv 1909.03725, arXiv.org E-Print Archive, available at: http://arxiv.org/abs/1909.03725, last access: 10 December 2019.

Jordan, A., Krüger, F., and Lerch, S.: Evaluating Probabilistic Forecasts with scoringRules, J. Stat. Softw., 90, 1–37, https://doi.org/10.18637/jss.v090.i12, 2019.

Junk, C., Monache, L. D., and Alessandrini, S.: Analog-Based Ensemble Model Output Statistics, Mon. Weather Rev., 143, 2909–2917, https://doi.org/10.1175/mwr-d-15-0095.1, 2015.

Klein, N., Kneib, T., Klasen, S., and Lang, S.: Bayesian Structured Additive Distributional Regression for Multivariate Responses, J. R. Stat. Soc. C-Appl., 64, 569–591, https://doi.org/10.1111/rssc.12090, 2014.

Lang, M. N., Mayr, G. J., Stauffer, R., and Zeileis, A.: Bivariate Gaussian models for wind vectors in a distributional regression framework, Adv. Stat. Clim. Meteorol. Oceanogr., 5, 115–132, https://doi.org/10.5194/ascmo-5-115-2019, 2019.

Lerch, S. and Baran, S.: Similarity-based semilocal estimation of post-processing models, J. R. Stat. Soc. C-Appl., 66, 29–51, https://doi.org/10.1111/rssc.12153, 2017.

Messner, J. W., Mayr, G. J., and Zeileis, A.: Heteroscedastic Censored and Truncated Regression with crch, R J., 8, 173–181, https://doi.org/10.32614/RJ-2016-012, 2016.

Möller, A., Spazzini, L., Kraus, D., Nagler, T., and Czado, C.: Vine Copula Based Post-Processing of Ensemble Forecasts for Temperature, arXiv 1811.02255, arXiv.org E-Print Archive, available at: http://arxiv.org/abs/1811.02255 (last access: 10 December 2019), 2018.

NASA JPL: NASA Shuttle Radar Topography Mission Global 30 Arc Second [Data Set], NASA EOSDIS Land Processes DAAC, https://doi.org/10.5067/MEaSUREs/SRTM/SRTMGL30.002, 2013.

Palmer, T. N.: The Economic Value of Ensemble Forecasts as a Tool for Risk Assessment: From Days to Decades, Q. J. Roy. Meteor. Soc., 128, 747–774, https://doi.org/10.1256/0035900021643593, 2002.

Pantillon, F., Lerch, S., Knippertz, P., and Corsmeier, U.: Forecasting Wind Gusts in Winter Storms Using a Calibrated Convection-Permitting Ensemble, Q. J. Roy. Meteor. Soc., 144, 1864–1881, https://doi.org/10.1002/qj.3380, 2018.

Rasp, S. and Lerch, S.: Neural Networks for Postprocessing Ensemble Weather Forecasts, Mon. Weather Rev., 146, 3885–3900, https://doi.org/10.1175/MWR-D-18-0187.1, 2018.

R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, available at: https://www.R-project.org/, last access: 10 December 2019.

Rodwell, M. J., Richardson, D. S., Parsons, D. B., and Wernli, H.: Flow-dependent reliability: A path to more skillful ensemble forecasts, B. Am. Meteorol. Soc., 99, 1015–1026, https://doi.org/10.1175/BAMS-D-17-0027.1, 2018.

Scheuerer, M.: Probabilistic Quantitative Precipitation Forecasting Using Ensemble Model Output Statistics, Q. J. Roy. Meteor. Soc., 140, 1086–1096, https://doi.org/10.1002/qj.2183, 2014.

Schlosser, L., Hothorn, T., Stauffer, R., and Zeileis, A.: Distributional Regression Forests for Probabilistic Precipitation Forecasting in Complex Terrain, Ann. Appl. Stat., 13, 1564–1589, https://doi.org/10.1214/19-AOAS1247, 2019.

Taillardat, M., Mestre, O., Zamo, M., and Naveau, P.: Calibrated Ensemble Forecasts Using Quantile Regression Forests and Ensemble Model Output Statistics, Mon. Weather Rev., 144, 2375–2393, https://doi.org/10.1175/mwr-d-15-0260.1, 2016.

Umlauf, N., Klein, N., and Zeileis, A.: BAMLSS: Bayesian Additive Models for Location, Scale, and Shape (and Beyond), J. Comput. Graph. Stat., 27, 612–627, https://doi.org/10.1080/10618600.2017.1407325, 2018.

Vogel, P., Knippertz, P., Fink, A. H., Schlueter, A., and Gneiting, T.: Skill of Global Raw and Postprocessed Ensemble Predictions of Rainfall over Northern Tropical Africa, Weather Forecast., 33, 369–388, https://doi.org/10.1175/waf-d-17-0127.1, 2018.

Wilson, L. J., Beauregard, S., Raftery, A. E., and Verret, R.: Calibrated Surface Temperature Forecasts from the Canadian Ensemble Prediction System Using Bayesian Model Averaging, Mon. Weather Rev., 135, 1364–1385, https://doi.org/10.1175/MWR3347.1, 2007.

Wood, S. N.: Generalized Additive Models: An Introduction with R, Chapman and Hall/CRC, https://doi.org/10.1201/9781315370279, 2017.

Article X

Lang M.N., Schlosser L., Hothorn T., Mayr G.J., Stauffer R., and Zeileis A. (2020). *Circular Regression Trees and Forests with an Application to Probabilistic Wind Direction Forecasting*. Journal of the Royal Statistical Society Series C: Applied Statistics, 69(5), 1357–1374, doi:10.1111/rssc.12437.

JCR ranking: **Category 1** in *Statistics and Probability*.

Contribution (CRT): *Conceptualization / data curation / software / validation / supervision / writing, review and editing*.

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2020)
69, *Part* 5, *pp.* 1357–1374

Circular regression trees and forests with an application to probabilistic wind direction forecasting

Moritz N. Lang and Lisa Schlosser,

Universität Innsbruck, Austria

Torsten Hothorn

Universität Zürich, Switzerland

and Georg J. Mayr, Reto Stauffer and Achim Zeileis

Universität Innsbruck, Austria

[Received January 2020. Revised July 2020]

Summary. Although circular data occur in a wide range of scientific fields, the methodology for distributional modelling and probabilistic forecasting of circular response variables is quite limited. Most of the existing methods are built on generalized linear and additive models, which are often challenging to optimize and interpret. Specifically, capturing abrupt changes or interactions is not straightforward but often relevant, e.g. for modelling wind directions subject to different wind regimes. Additionally, automatic covariate selection is desirable when many predictor variables are available, as is often the case in weather forecasting. To address these challenges we suggest a general distributional approach using regression trees and random forests to obtain probabilistic forecasts for circular responses. Using trees simplifies model estimation as covariates are used only for partitioning the data and subsequently just a simple von Mises distribution is fitted in the resulting subgroups. Circular regression trees are straightforward to interpret, can capture non-linear effects and interactions, and automatically select covariates affecting location and/or scale in the von Mises distribution. Circular random forests regularize and smooth the effects from an ensemble of trees. The new methods are applied to probabilistic wind direction forecasting at two Austrian airports, considering other common approaches as a benchmark.

Keywords: Circular data; Distributional regression; Probabilistic forecasting; Random forests; Regression trees; von Mises distribution

1. Introduction

Circular data can be found in a variety of applications and subject areas, e.g. hourly crime rate in socio-economics, animal movement direction or gene structure in biology, and wind direction as one of the most important weather variables in meteorology. Fitting a statistical model to this type of data requires the incorporation of its specific feature of periodicity. For example, angular data are restricted to an interval such as $[0, 2\pi)$ with 0 being equivalent to 2π.

Address for correspondence: Moritz N. Lang, Department of Statistics, Faculty of Economics and Statistics, Universität Innsbruck, Universitätsstrasse 15, 6020 Innsbruck, Austria.
E-mail: Moritz.Lang@uibk.ac.at

1.1. Circular regression: conditional mean versus distributional models

Many approaches to model circular data assume that the circular variable of interest follows a circular distribution, in particular the von Mises distribution which is also known as the 'circular normal distribution'. One of the first regression models with a circular response variable and linear covariates was presented by Gould (1969) where the circular mean was predicted by a linear combination of covariates. Johnson and Wehrly (1978) refined this idea by plugging in a link function transforming the linear predictor to a restricted interval of length 2π. This generalized linear model (GLM) type of approach was further extended by Fisher and Lee (1992) and subsequently by Fisher (1993) introducing independent GLMs for either the location or scale parameter with appropriate link functions while keeping the other parameter constant. Additionally, they developed a combined heteroscedastic or distributional version with alternating re-estimation of the location and scale parameters conditionally on the respective sets of covariates until convergence (Fisher and Lee, 1992; Fisher, 1993). Although all of these models are built on well-elaborated theory, their application in practice remains very challenging, mainly because of the complexity that is encountered in optimizing the corresponding log-likelihood function which is not globally concave. Therefore, highly informative starting values are crucial for such circular GLMs to converge (Pewsey *et al.*, 2013; Gill and Hangartner, 2010). To avoid this strong dependence on appropriate initial values, Mulder and Klugkist (2017) presented a Bayesian alternative of a homoscedastic GLM for circular data. However, apart from potential difficulties in the optimization procedure of circular GLMs, the interpretation of the underlying additive effects is often challenging as well because the link function is highly non-linear and the representation of smooth transitions on the unit circle is not straightforward. For example, the same rotation can be obtained in either positive or negative direction on the circle, leading to an ambiguous interpretation.

As a very intuitive and data-driven alternative, we propose a flexible tree-based regression approach for modelling circular data by applying the von Mises distribution within the methodology of distributional trees and forests (Schlosser *et al.*, 2019a). The resulting circular regression trees and forests avoid the difficulties discussed of circular GLMs by using the available covariates for partitioning the data into sufficiently homogeneous subgroups so that a simple von Mises distribution without further covariates can be fitted to the circular response in each of these subgroups. This obviates the need for a link function or for iterating between models for the separate distribution parameters. By leveraging the distributional modelling approach, the trees can automatically detect and capture differences in both distribution parameters, providing a fully specified circular response distribution in each terminal node, offering a wide range of statistical inference. In addition, the tree structure employed can capture non-additive effects whereas forests enable the modelling of smooth changes. Furthermore, covariates and their possible interactions do not need to be specified in advance as they are selected automatically in the recursive partitioning algorithm.

This novel approach to circular regression trees and forests complements the literature on tree-based circular modelling. Lund (2002) has already introduced a circular regression tree algorithm where binary splits are made based on an angular distance measure capturing node homogeneity. However, this only models changes in the conditional mean, whereas modelling the conditional variance or full probabilistic distribution would also allow uncertainty in the forecast to be estimated (Gneiting, 2008). Moreover, Hara and Chellappa (2017) introduced both regression trees and random forests for image-based object direction estimation. These extend classical conditional mean regression trees (Breiman *et al.*, 1984) and corresponding random forests (Breiman, 2001) to conditional mean models for circular data. But, rather than considering only binary splits, they employed k-means clustering to determine multiway splits

with the number k adaptively determined based on the Bayes information criterion with either a Gaussian or von Mises kernel. Although a distributional aspect is considered in this step of the algorithm, variances are always restricted to be shared across splits and, more importantly, no probabilistic forecasts are obtained. Therefore, our proposed circular regression trees and random forests take these ideas a step further: a fully distributional approach is adopted for all aspects of the algorithms from selection of the split variables, selection of the split points and ensemble aggregation of the predictions across trees in a forest. This provides a full probabilistic predictive model, as proposed by Gneiting (2008).

1.2. Motivating example

To provide a first impression of the methodology presented, a circular regression tree is employed for probabilistic wind direction forecasting. Wind direction is a classical circular quantity and accurate forecasts are of great importance for decision-making processes, e.g. in air traffic management as considered in this study. Fig. 1 shows an estimated tree for 1-hourly forecasts at Innsbruck Airport, which is at the bottom of a narrow valley within the European Alps. Topography channels the wind along the west–east valley axis or along a tributary valley intersecting from the south. Hence, pressure gradients to which valley wind regimes react are considered as covariates along with other meteorological measurements (lagged by 1 h) and their derivatives, such as wind direction and wind speed at the airport itself as well as spatial and temporal differences.

Fig. 1 illustrates the resulting tree along with the empirical (grey) and fitted von Mises (red) wind direction distribution in each terminal node. Based on the fitted location parameters $\hat{\mu}$, the subgroups can be distinguished into the following wind regimes:

(a) up-valley winds blowing from the valley mouth towards the upper valley (from east to west; nodes 4 and 5);
(b) downslope winds blowing across the Alpine crest along the intersecting valley towards Innsbruck (from south–east to north–west; node 8);
(c) down-valley winds blowing in the direction of the valley mouth (from west to east; nodes 10, 12 and 13).

Node 7 captures observations with quite low wind speeds that cannot be clearly distinguished into specific wind regimes and are consequently associated with a very low estimated concentration parameter $\hat{\kappa}$, i.e. a high estimated variance. In terms of covariates, the lagged wind direction ('persistence') is mostly responsible for distinguishing the broad range of wind regimes that were listed above whereas the pressure gradients and wind speed separate the data into subgroups with high *versus* low precision. A more comprehensive application of circular regression trees and forests to probabilistic wind direction forecasting is presented in Section 4 for Innsbruck Airport and Vienna International Airport, together with a comparison with commonly used alternative approaches.

The remainder of the paper is structured as follows. The theory on probabilistic circular modelling, introducing the von Mises distribution, and circular regression models are discussed in Section 2. The methodology of circular regression trees and forests and their features are introduced in Section 3. After the application presented in Section 4, a comprehensive summary and conclusions are given in Section 5.

2. Probabilistic circular modelling

Probabilistic modelling of circular data requires the selection of a probability distribution which

Fig. 1. Fitted tree based on the von Mises distribution for 1-hourly wind direction forecasts at Innsbruck Airport: in each terminal node the empirical histogram (■) and fitted density (⬭) are depicted along with the estimated location parameter (—); the covariates selected for splitting are wind direction (meteorological degree), wind speed (metres per second) and pressure gradients (dpressure; hectopascals) west, east and south of the airport, all lagged by 1 h; in the meteorological context wind direction is defined on the scale [0°, 360°] and increases clockwise from north (0°)

accounts for the periodicity of circular data. Generally, this feature can be obtained by 'wrapping' the probability density function of any continuous distribution around the unit circle (Mardia and Jupp, 1999). In that way, the wrapped Cauchy distribution or the wrapped normal distribution can be employed to model symmetric unimodal circular data. A close approximation to the wrapped normal distribution that is mathematically simpler and hence easier to use is provided by the von Mises distribution (Fisher, 1993), which is a purely circular distribution that is also known as 'the circular normal distribution' and is a common choice for probabilistic modelling of circular data. Based on a location parameter $\mu \in [0, 2\pi)$ and a concentration parameter $\kappa > 0$ the density of the von Mises distribution for an observation $y \in [0, 2\pi)$ is given by

$$f_{\mathrm{vM}}(y; \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} \exp\{\kappa \cos(y - \mu)\}, \tag{2.1}$$

where $I_0(\kappa)$ is the modified Bessel function of the first kind and order 0 (see, for example, Mardia and Zemroch (1975), Jammalamadaka and Sengupta (2001) or Ley and Verdebout (2017) for a more detailed overview).

The corresponding log-likelihood function is defined as

$$l(\mu, \kappa; y) = \log\{f_{\mathrm{vM}}(y; \mu, \kappa)\} = -\log\{2\pi I_0(\kappa)\} + \kappa \cos(y - \mu). \tag{2.2}$$

To fit a probabilistic model $\mathrm{vM}(y; \mu, \kappa)$ to a circular response y, the distribution parameters μ and κ need to be estimated. This can be done by maximizing the log-likelihood function $l(\mu, \kappa; y)$:

$$(\hat{\mu}, \hat{\kappa}) = \underset{\mu, \kappa}{\arg\max} \sum_{i=1}^{n} l(\mu, \kappa; y_i) \tag{2.3}$$

yielding maximum likelihood estimators $\hat{\mu}$ and $\hat{\kappa}$ such that a fully specified distributional model is fitted to the learning data $\{y_i\}_{i=1,\ldots,n}$.

The score function

$$\begin{aligned} s(\mu, \kappa, y) &= \left(\frac{\partial l}{\partial \mu}(\mu, \kappa; y), \ \frac{\partial l}{\partial \kappa}(\mu, \kappa; y) \right) \\ &= \left(\kappa \sin(y - \mu), \ -\frac{I_1(\kappa)}{2\pi I_0(\kappa)} + \cos(y - \mu) \right) \end{aligned} \tag{2.4}$$

provides a way to obtain a measure of goodness of fit of the model for each observation and fitted parameter. Then, the optimization problem in equation (2.3) can alternatively be specified as

$$\sum_{i=1}^{n} s(\hat{\mu}, \hat{\kappa}, y_i) = 0. \tag{2.5}$$

Fig. 2 depicts a von Mises model for circular data in $[0, 2\pi)$ fitted by maximum likelihood, either by using a linearized (Fig. 2(a)) or circular (Fig. 2(b)) scale. In both cases, the empirical histogram (grey bars) is shown along with the fitted density (the red line) and estimated location parameter (the red hand). However, this distributional model considers only the circular response variable but no covariate. Of course, including covariates is of interest in a regression set-up for forecasting.

In most GLM or generalized additive model approaches to circular regression the location parameter μ depends on covariates $\mathbf{z}$ through a link function $g(\cdot)$, circular intercept μ_0 and coefficient vector $\boldsymbol{\beta}$:

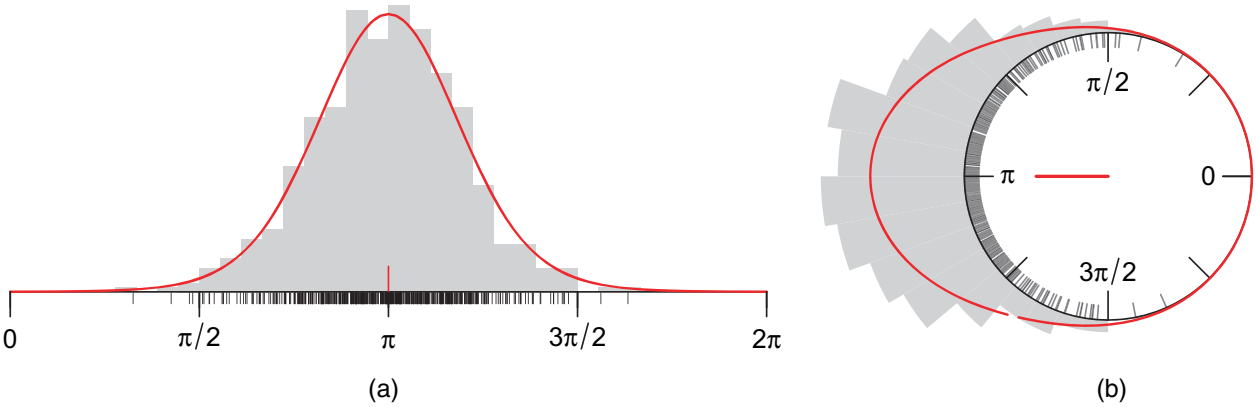

Fig. 2. Illustration of a von Mises model for circular data in the interval $[0, 2\pi)$ fitted by maximum likelihood (in both panels the empirical histogram (■) and fitted density (⬭) are depicted along with the estimated location parameter (—)): (a) linearized scale; (b) circular scale

$$\mu = \mu_0 + g(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}). \tag{2.6}$$

The link function transforms the additive predictor to an interval of length 2π. Typically $g(x) = 2\tan^{-1}(x)$ is employed, as suggested by Fisher and Lee (1992). They also developed a heteroscedastic version by combining two individual GLMs, each for one of the parameters μ and κ. This provides a first approach to a fully probabilistic regression model for circular data, albeit the parameters are not regressed simultaneously on covariates as in the more general framework that is provided by generalized additive models for location, scale and shape (Rigby and Stasinopoulos, 2005). Nevertheless, other circular additive models share the previously discussed difficulties that are induced by the characteristics of the log-likelihood function and the strongly non-linear link function. In general, for additive models a proper model specification can be very challenging, particularly for a high number of covariates and no information on possible interactions. Moreover, the additive structure might impose a smooth effect even if the true underlying effect is an abrupt shift, which occurs, for example, in atmospheric wind fields.

In contrast, the tree-based circular regression models that are proposed in the next sections largely avoid the problems above by employing recursive partitioning in combination with local adaptive likelihood estimation.

3. Circular regression trees and forests

Starting out from the ideas of Lund (2002), we introduce circular regression trees and forests considering splits in all distribution parameters of the von Mises distribution and providing a full probabilistic model. Moreover, the resulting tree-based models provide a very intuitive and data-driven alternative to commonly used GLMs for circular data.

3.1. Circular regression trees

Fitting a global model to a full data set can be very challenging, particularly for complex data with substantial variations. Therefore, separating the data set into more homogeneous subgroups based on covariates before fitting a local model in each of these subgroups enables (potential) group-specific effects to be captured more precisely and hence can result in an overall better-fitting model. This is the general idea of regression trees, which were combined with distributional modelling in Schlosser *et al.* (2019a). Specifying a full distributional model in each node of the tree yields a distributional regression tree, where selecting the von Mises

distribution enables an application to circular data. The crucial step of how and where to split the data can be accomplished with the unbiased recursive partitioning algorithms MOB (Zeileis *et al.*, 2008) or CTree (Hothorn *et al.*, 2006). For this, model scores are obtained by evaluating the score function $s(\cdot)$ for each individual observation at the parameter estimates (equation (2.4)). For the von Mises distribution with its two distribution parameters (μ and κ) and a data set of n observations, this yields an $n \times 2$ matrix that can be employed as a discrepancy measure, capturing how well each given observation conforms with the estimated location $\hat{\mu}$ and precision $\hat{\kappa}$ respectively. To capture dependence on covariates, the association between the model's scores and each available covariate is assessed by using either a parameter instability test (MOB) or a permutation test (CTree). By doing so in each partitioning step, the covariate with the highest significant association (i.e. lowest significant p-value, if any) is selected for splitting the data. The corresponding split point is chosen either by optimizing the log-likelihood (MOB) or a two-sample test statistic (CTree) over all possible partitions. This procedure is repeated recursively until there are no significant parameter instabilities or until another stopping criterion has been met (e.g. subgroup size or tree depth). A more detailed description of the applied tree building algorithm can be found in Appendix A.

Once a distributional tree model has been fitted it can be employed to obtain probabilistic predictions for a possibly new set of observed covariates $\mathbf{z} = (z_1, \ldots, z_m)$. Starting at the root node, the tree structure leads the observation to a terminal node where the parameter pair $(\hat{\mu}, \hat{\kappa})$ is estimated for the corresponding subset of learning observations. This can also be expressed by employing weights which indicate whether the ith learning observation and the observation $\mathbf{z}$ belong to the same terminal node:

$$w_i^{\mathrm{T}}(\mathbf{z}) = \sum_{b=1}^{B} \mathbf{1}\{(\mathbf{z}_i \in \mathcal{B}_b) \wedge (\mathbf{z} \in \mathcal{B}_b)\}. \tag{3.1}$$

Here, $\mathbf{1}(\cdot)$ is the indicator function and $\mathcal{B}_b$ is the bth out of B segments partitioning the covariate space in disjoint subsets. Then the estimated parameter pair $(\hat{\mu}, \hat{\kappa})(\mathbf{z})$ specifying the predicted von Mises distribution for a given $\mathbf{z}$ is obtained by a weighted maximum likelihood estimator:

$$(\hat{\mu}, \hat{\kappa})(\mathbf{z}) = \underset{\mu,\kappa}{\arg\max} \sum_{i=1}^{n} w_i^{\mathrm{T}}(\mathbf{z})\, l(\mu, \kappa; y_i). \tag{3.2}$$

Therefore, the same parameter pair is estimated for all observations belonging to the same terminal node, which speeds up computation since the parameter estimates do not need to be recalculated for each (new) observation via maximum likelihood but can be extracted directly from the learning sample and the fitted model.

Whereas tree models can capture non-additive effects, their structure and the consequential strict separation of data into subgroups hinder an adequate depiction of smooth effects. They can be included by combining an ensemble of trees to obtain a regression forest, which also stabilizes the model.

3.2. Circular regression forests

Ensembles or forests are a natural extension of (circular) regression trees that can improve forecasts by regularizing and stabilizing the model. Random forests introduced by Breiman (2001) average the predictions of an ensemble of trees, each built on a subsample or bootstrap of the original data. A generalization of this strategy is to obtain weighted predictions by adaptive local likelihood estimation of the distributional parameters (section 2.3 of Schlosser *et al.* (2019a); Hothorn and Zeileis (2017)). More specifically, for each (possibly new) observation

$\mathbf{z}$ a set of averaged 'nearest neighbour' weights $w_i^{\mathrm{F}}(\mathbf{z})$ is obtained that is based on the number of trees in which $\mathbf{z}$ is assigned to the same terminal node as each learning observation y_i, $i \in \{1, \ldots, n\}$. Hence, for a forest of T trees, the weights are calculated via

$$w_i^{\mathrm{F}}(\mathbf{z}) = \frac{1}{T} \sum_{t=1}^{T} \sum_{b=1}^{B^t} \frac{\mathbf{1}\{(\mathbf{z}_i \in \mathcal{B}_b^t) \wedge (\mathbf{z} \in \mathcal{B}_b^t)\}}{|\mathcal{B}_b^t|}, \tag{3.3}$$

where $|\mathcal{B}_b^t|$ denotes the number of observations in the bth segment of the tth tree. Therefore, similar observations ending up more often in the same terminal node have higher weights and thus a stronger influence in the weighted maximum likelihood estimation.

In that way a specific set of weights can be calculated for each observation, yielding its specific parameter estimates for the von Mises distribution

$$(\hat{\mu}, \hat{\kappa})(\mathbf{z}) = \underset{\mu,\kappa}{\arg\max} \sum_{i=1}^{n} w_i^{\mathrm{F}}(\mathbf{z})\, l(\mu, \kappa; y_i). \tag{3.4}$$

Therefore, the resulting parameter estimates can smoothly adapt to the given covariates $\mathbf{z}$ whereas $w_i(\mathbf{z}) = 1$ would correspond to the unweighted full sample estimates and $w_i(\mathbf{z}) \in \{0, 1\}$ corresponds to the subgroup selection from a tree. Thus, circular regression forests can capture both smooth and abrupt changes, whereas covariates and possible interactions are selected automatically and do not explicitly need to be specified beforehand.

4. Application: probabilistic wind direction forecasting

As motivated in Section 1, accurate forecasts of wind directions are of great importance for risk management in various fields such as agriculture, energy production or aviation. For example, to direct airplanes to a safe landing, precise knowledge of wind direction for the next hour(s) at the airport is highly desirable and adequate prediction methods are required. This section exemplifies the use of circular regression trees and forests with wind direction forecasts for two Austrian airports—one in flat terrain; the other in mountainous terrain. The results are benchmarked against alternative probabilistic forecasting methods. The application is based on 1-h and 3-h forecasts employing lagged observations in the vicinity of the airports as possible predictor variables.

4.1. Data

The circular response variable that is considered in this application is a 10-min average of wind direction measurements at Innsbruck Airport and Vienna International Airport on an hourly temporal resolution. Temporal information and 1-hourly resolved 10-min mean observations of various meteorological quantities are used as predictor variables, including wind direction, wind speed, temperature, air pressure and humidity, all lagged by 1 h or 3 h according to the respective forecasting step. The meteorological variables are measured either directly at the airports or within their vicinities. For Innsbruck, measurements at the airport and along the intersecting valleys are used, whereas, for Vienna, measurements at the airport and within its vicinity of approximately 30 km are used. Fig. 3 provides a topographical overview of the airports and their surrounding areas with the station sites that are employed in the application. In addition, we use derived quantities such as 3-h means, minima and maxima, as well as 1- and 3-h temporal changes and spatial differences towards the airport of the respective quantities. An overview of the data sets employed can be found in Table 1.

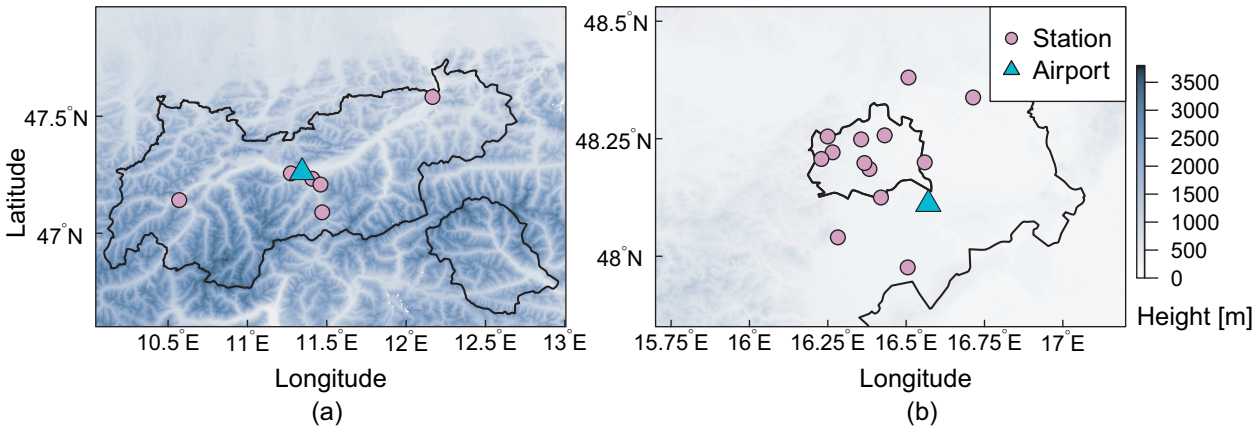

Fig. 3. Overview of the study area for (a) Innsbruck Airport and (b) Vienna International Airport: for Innsbruck, four stations at the airport and six stations along the intersecting valleys are used, whereas, for Vienna, nine stations at the airport and 13 stations within the vicinity of approximately 30 km are used; elevation data are obtained from the TandDEM-X digital elevation model with a horizontal resolution of 90 m (Wessel *et al.*, 2018)

Table 1. Overview of the data sets employed in the application to probabilistic wind direction forecasting†

Data components	*Description*
Temporal information	Time of the day, day of the year
Meteorological variables	Wind direction, wind (gust) speed, (reduced) air pressure, relative humidity, temperature
Derived quantities	3-hourly means, minima and maxima, 1-hourly and 3-hourly temporal changes, spatial differences towards the airport
Weather stations (Innsbruck)	4 stations at the airport, as well as Igls, Kematen, Kufstein, Landeck, Patscherkofel and Steinach
Weather stations (Vienna)	9 stations at the airport, as well as Arsenal, Donaufeld, Exelberg, Gänserndorf, Groß-Enzersdorf, Gumpoldskirchen, Hohe Warte, Innere Stadt, Jubiläumswarte, Mariabrunn, Seibersdorf, Unterlaa and Wolkersdorf

†For Innsbruck and Vienna, various meteorological variables and derived quantities of these are considered at the respective stations, located either directly at the airports or in their vicinities.

The data that are used in the application consist of five subsequent years from January 2014 to December 2018. After first eliminating predictor variables with more than 5% missing values and then time points with any missing observations, the data set consists of 41979 time points and 260 covariates for Innsbruck, and of 38985 time points and 494 covariates for Vienna.

4.2. Models and evaluation

For a fair evaluation of circular regression trees and forests, and to investigate whether they can be applied as a reasonable alternative to already existing approaches, three additional statistical models are employed in the application to probabilistic wind direction forecasting. Two of them are based on existing approaches used in the meteorological field, whereas the third is a state of the art GLM-type model to forecast circular response variables.

(a) *Climatological model*: accurate knowledge of weather quantities' climatologies can be important for a wide range of applications. Although forecasts based on climatologies,

by construction, do not adapt to the current weather situation they are still a useful baseline for the validation of newly developed forecasting systems (Simon *et al.*, 2017; Stauffer *et al.*, 2017).

Specifically, the climatology that is employed in what follows uses all observations at the same time (to adapt to daily cycles) in a window of 31 days centred on the day of interest (to adapt to seasonal cycles) in all available years in the sample. Based on these observations a probabilistic model is obtained by maximum likelihood estimation as described in Section 2. This approach follows Vogel *et al.* (2018) and has been discussed in a comprehensive summary on different time-adaptive training schemes in Lang *et al.* (2020).

(b) *Persistency model*: the persistence describes the previous value of a single weather quantity in a time series. Like the climatology it is a very basic prediction model that is often applied as a baseline reference in weather forecasts (National Oceanic and Atmospheric Administration National Weather Service, 2019). Especially in nowcasting tasks with very short forecasting steps, the persistence can provide very good estimates.

To obtain a full probabilistic persistency model, we proceed similarly to the climatological model by using maximum likelihood estimation and fitting the distribution parameters of the von Mises distribution conditional on lagged response values according to the description in Section 2. We fit one model for every hour throughout the validation data set employing the previous six lagged response values as training data. To allow for a stronger influence of observations closer to the time of interest, exponential smoothing is employed with a smoothing factor of 0.5; accordingly, for every prediction an equal influence rate of 50% is assigned both to the current observation and to the previous five observations together. Observations with longer time lags have exponential weights below 0.01 and are therefore omitted from the training data.

(c) *GLM*: traditional approaches to forecast circular response variables are often based on circular GLM-type models (Fisher, 1993). As discussed in Section 1, circular regression models often experience the problem that the likelihood function can be strongly irregular, which makes optimization rather difficult. Hence, they often do not converge if no appropriate initial values are provided (Pewsey *et al.*, 2013; Gill and Hangartner, 2010). In this study, to be able to employ a GLM without further parameter tuning as a reference, we use the Bayesian implementation of Mulder and Klugkist (2017) which depends less on initial values because of a Markov chain Monte Carlo sampling algorithm using weakly informative priors.

Following Fisher and Lee (1992) the model uses a link function $g(\cdot)$ to keep the response values within an interval of length 2π. Because in the implementation of Mulder and Klugkist (2017) the accommodation of circular covariates would require an additional manual adjustment, such as an approximation by a finite Fourier expansion (see Lund (1999)), we use only the linear components of the lagged two-dimensional wind vector $(u, v)^{\mathrm{T}}$ and the lagged wind speed spd as predictor variables. Thus, the model formula for the location parameter μ of the von Mises distribution can be written as

$$\mu = \beta_0 + g(\beta_1 u + \beta_2 v + \beta_3 \mathrm{spd}) \tag{4.1}$$

with β_0 being a circular intercept, $\beta_\cdot$ the regression coefficients and the link function $g(x) = 2\tan^{-1}(x)$. In addition, a constant concentration parameter κ is fitted to the full learning sample. To allow for seasonally varying error characteristics in both bias and slope coefficients, and to allow for seasonal heteroscedasticity that is captured by the concentration parameter, we use the same time-adaptive training approach as for the climatological model; hence, separate models are estimated over all observation dates,

using the same time of 31 days centred on the day of interest over all available years in the training data (Lang *et al.*, 2020).

(d) *Circular regression tree*: for the circular regression tree that was introduced in Section 3.1, all covariates that are provided in the learning data can be considered because of an intrinsic automatic variable selection performed in the tree estimation. The tree is built with the newly developed R package `circtree` employing the CTree algorithm (Hothorn *et al.*, 2006) using a minimal number of 2000 observations in each terminal node (argument `minbucket`).

(e) *Circular regression forest*: following the description in Section 3.2, the circular regression forest that is used in this study is constructed based on 100 individual trees employing the R package `circtree`. Each of these trees is again built by the CTree algorithm on a subsample containing 30% of the original learning data. All covariates are included for building each tree which ensures that the lagged response variable is always considered for splitting. This bagging approach can be applied in `circtree` by setting the argument `mtry` to the total number of covariates. Since a high number of possible split points leads to high computational costs, the covariates are binned in a maximum of 50 classes (argument `nmax=c(yx=Inf,z=50)`). Contrary to a single-tree model, forests usually consist of very large trees as they are not prone to overfitting the data because of the stabilization that is obtained by combining the individual trees. Therefore, we use the following control arguments to build rather large trees: the minimal number of observations to perform a split is set to 20 (argument `minsplit`), the minimal number of observations in each terminal node is set to 7 (argument `minbucket`), and the level of significance for variable selection is kept at its maximum value of 1 (argument `alpha`).

To compare the predictive performance of all the models proposed, a circular analogue of the continuous ranked probability score (CRPS) as introduced by Grimit *et al.* (2006) is computed. Just like the linear version of the CRPS (for more details see Hersbach (2000)) it is a proper scoring rule (Gneiting and Raftery, 2007) and measures the difference between an observation and the corresponding predicted distribution function to assess the probabilistic goodness of fit for the estimated model. Hence, the lower the CRPS value the better is the predictive performance. Contrary to the linear version, the circular CRPS reduces not to the absolute error but to the angular distance when the forecast is deterministic.

In addition to the raw CRPS, corresponding skill scores are computed to assess differences in the improvement of the various statistical models over the climatological model that is used as a reference:

$$\mathrm{CRPSS}(\mathrm{model}) = 1 - \frac{\mathrm{CRPS}(\mathrm{model})}{\mathrm{CRPS}(\mathrm{climatology})}. \tag{4.2}$$

All scores that are presented in the next section are computed out of sample based on 5 years of data. For the persistency model only dates before the time of interest are used and the validation is performed rolling over all observations. For all other models fivefold cross-validation is employed using up to four calendar years for model training and the remaining single calendar year for validation. Because of the large sample size of 24 hourly values per day over 5 years, some kind of temporal aggregation is needed to ensure a correct visual comparison of the individual methods. The analyses that were performed have shown that for the models employed the variability of the predictive performance over the 5 years is lower than over a single day or over a single year. Hence, CRPS and CRPS skill scores are aggregated over the respective five validation years which yields 24 hourly scores per month averaged over the five validation years.

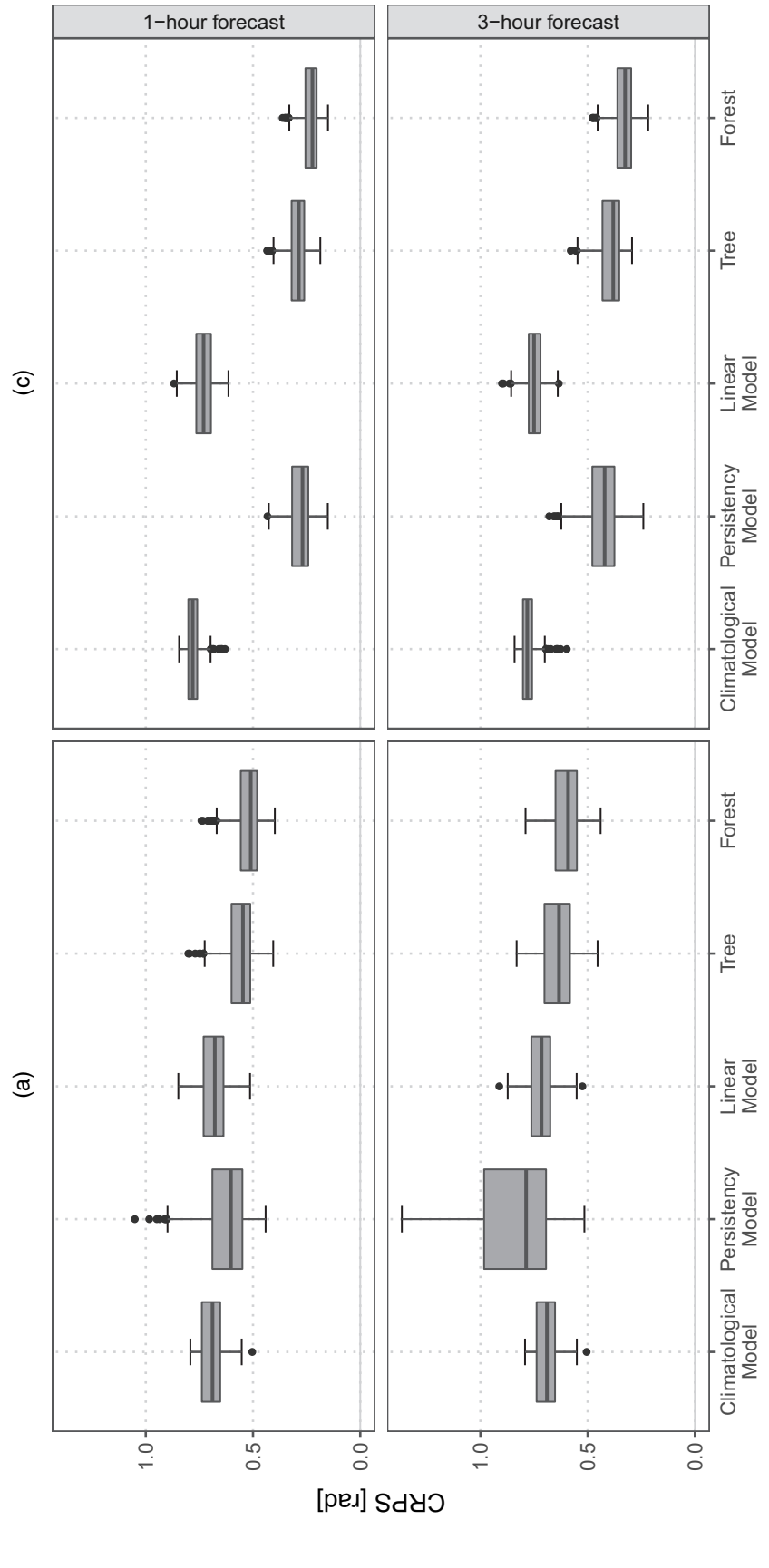

Fig. 4. CRPS skill scores of wind direction forecasts based on the full predictive von Mises distribution for 1-h and 3-h forecasts at (a), (b) Innsbruck Airport and (c), (d) Vienna International Airport: each box-and-whisker contains 24 hourly scores for each of the 12 months averaged over the five validation years, which yields a total of 288 yearly mean values; the scores are shown for the climatological, persistency and the linear model as well as for the circular regression tree and forest

4.3. Results

This section provides a detailed analysis on the predictive performance of the different proposed statistical models applied to probabilistic wind direction forecasting. To ensure a comprehensive comparison of the models, wind direction forecasts are evaluated for two different lead times at two airports with different climatological site characteristics. Fig. 4 shows the CRPS values of the employed models at forecast steps 1 h and 3 h for Innsbruck (Figs 4(a) and 4(b)) and Vienna (Figs 4(c) and 4(d)). The scores are aggregated over the five validation years, yielding yearly mean values for every hour per calendar month, with a lower score indicating better performance. The circular regression forest overall provides the best predictive performance, followed by the circular regression tree and the persistency model for both stations at both forecast steps, except for the 3-h forecasts at Innsbruck where the persistency model is outperformed by all the others. In comparison with the circular regression tree and forest, for both stations and forecast steps, the climatological model and the linear model show clearly higher CRPS values and hence a lower predictive performance. The different site characteristics of the airports Innsbruck (Figs 4(a) and 4(b)) and Vienna (Figs 4(c) and 4(d)) seem to have an effect on the absolute level of the model performances and on their respective predictive performance variances. At Innsbruck, because of the surrounding mountains only a limited number of possible wind directions exists, namely the three wind regimes that were discussed for Fig. 1 in Section 1. Therefore, for Innsbruck the wind direction remains quite constant in one of these possible states, but once a change takes place it is mostly a major wind direction shift, e.g. from up valley to down valley. Because of the few wind regimes the rather inflexible climatological and linear models score relatively well with similar CRPS values to those of the other models (Figs 4(a) and 4(b)). In addition, at Innsbruck the potentially high prediction errors in case of a change in the wind regime seem to lead to a higher variation in the predictive performance for all models in comparison with Vienna; this variation is especially high for the persistency model because of its strong vulnerability to abrupt wind shifts. In contrast, at Vienna smaller and less abrupt changes in the wind direction as well as less pronounced wind regimes are observed due to the less mountainous surrounding. This seems to weaken the predictive performance of the climatological and linear models, and to reduce the performance variability for all models (Figs 4(c) and 4(d)).

The different forecast steps have apparently only a minor effect on the predictive performance of the climatological model and the linear model at both stations. As expected, for the persistency model, at both stations, higher scores for the 3-h forecast (Figs 4(b) and 4(d)) reveal a lower performance for longer lead times; this is due to the lower information content of 3-hourly instead of 1-hourly lagged response values employed as covariates in the persistency model. The circular regression tree and forest seem to compensate partially for the lower skill of the lagged response values by other covariates; hence their predictive performance only slightly decreases for the longer lead time. This compensation is especially evident for Innsbruck, where the difference in performance between the persistency model and the tree-based methods significantly increases from the 1-hourly to the 3-hourly forecast.

In addition to the raw CRPS (Fig. 4), CRPS skill scores with the climatological model as a reference are provided in Fig. 5. Skill scores are in per cent, where positive values indicate an improvement in the predictive performance over the reference. For all set-ups, the circular regression forest has the highest skill scores with a mean performance gain of 13–25% and 58–71% for Innsbruck and Vienna respectively. As discussed for Fig. 4, this improvement over the climatological model is lower for Innsbruck because of the low number of predominant wind regimes and hence a relatively good performance of the climatological model. Additionally, Fig. 5 shows that although the persistency model's performance is lower than the reference (Fig. 5(b)) the tree-based models can compensate for the low skill of the lagged

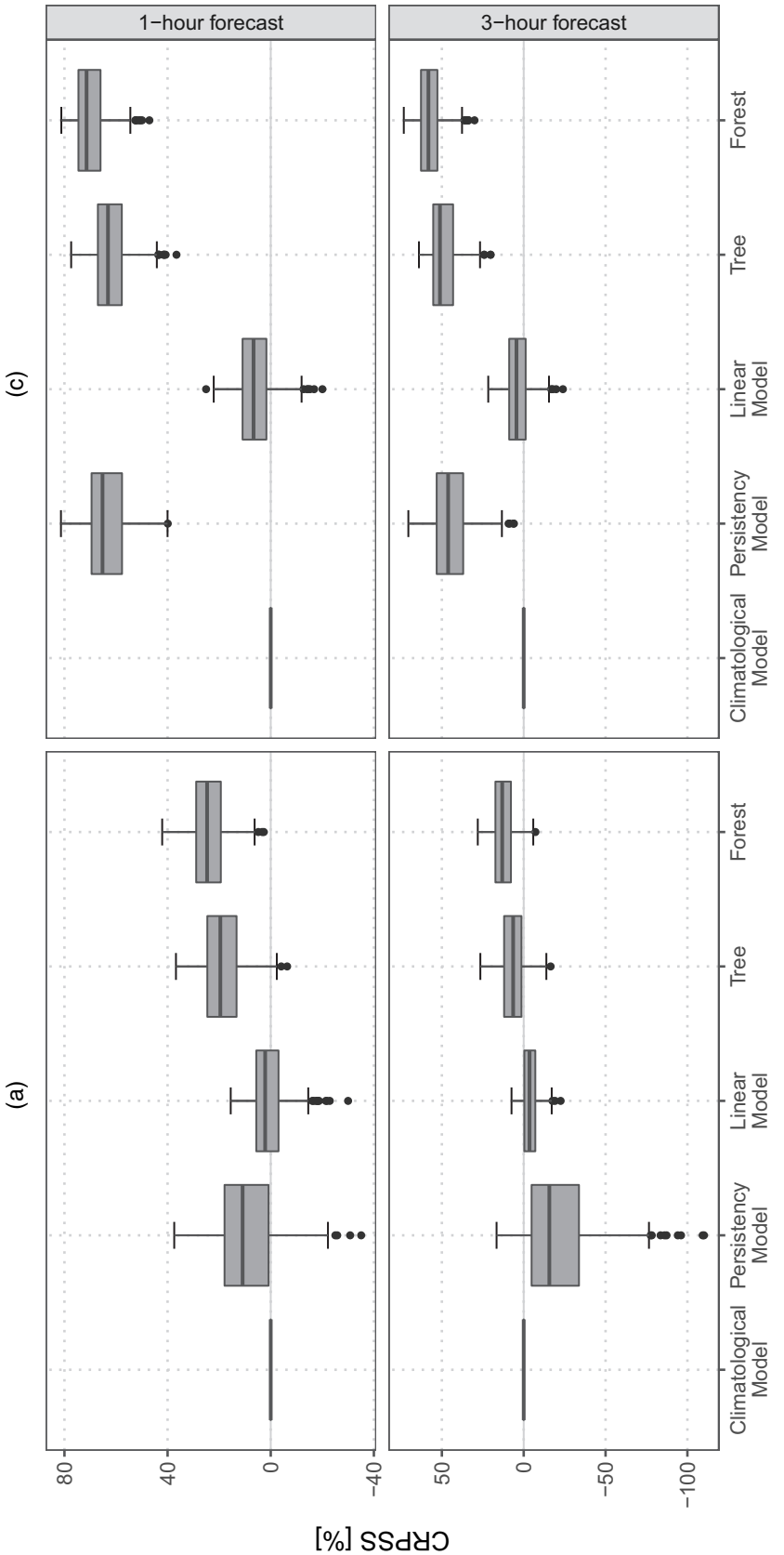

Fig. 5. As for Fig. 4, but showing CRPS skill scores with the climatological model as reference (skill scores are in per cent; positive values indicate improvements over the reference): (a), (b) Innsbruck Airport; (c), (d) Vienna International Airport

response values that are employed as covariates and, hence, are still significantly superior to the reference.

5. Summary and conclusion

Extending the toolbox for modelling circular data, circular regression trees and forests are established by coupling model-based recursive partitioning with the von Mises distribution. By separating the data into more homogeneous subgroups, possible difficulties in circular regression are bypassed as covariates are solely considered for splitting and group-specific models are fitted without further covariates. In addition, by specifying the von Mises distribution for each node and allowing for splits in both distribution parameters μ and κ, fully probabilistic forecasts are provided.

The performance of the novel circular regression trees and forests is assessed in an application on short-term probabilistic wind direction forecasting at two airports with different site characteristics. As benchmark models, probabilistic climatology and persistency models, as well as a state of the art circular GLM-type model, are evaluated based on proper scoring rules. In summary, the circular regression trees and forests have the highest predictive performance in this setting. For cases without changes in the wind regime, lagged response values already provide highly skilful estimates, leading to a good performance of the persistency model as observed for short-term wind direction forecasts in the application. While in these cases the trees and forests also benefit from the highly informative lagged response, they can compensate for lower information of this covariate by incorporating other quantities and possible interactions of these, in contrast with the persistency model (see Fig. 5). Hence, the tree-based models provide reliable forecasts in all the meteorological settings tested. For operational use, a possible extension could be the incorporation of numerical weather predictions as (additional) covariates. Although this probably only slightly improves the predictive skill for short lead times, it possibly extends the potential forecast range of the various methods.

For the specific task of wind forecasting, the wind direction is often only relevant if the wind speed is sufficiently high. Hence, it is of interest to account for both quantities simultaneously, e.g. by considering a bivariate normal distribution for wind vectors (from which wind speed and wind direction can be obtained). The parameters of this bivariate normal distribution could then be linked to available covariates by using an additive regression framework (as proposed by Lang *et al.* (2019)) or using a tree-based approach, similar to that proposed in this paper. Moreover, a rather different approach for a combined response of wind speed and wind direction would be a two-step or hurdle model: in the first step this could build on the truncated normal model of Thorarinsdottir and Gneiting (2010) to capture wind speed; in the second step a circular wind direction model is used given that a certain hurdle for the wind speed is crossed.

Another possible improvement for obtaining more parsimonious circular regression trees is to consider splitting circular covariates into two circle segments by searching two split points simultaneously rather than sequentially at different depths. Although this might slightly improve the predictive performance of circular regression trees, this should not affect the performance of the forests, as they consist of very large trees with many different splits.

To conclude, in general the tree structure can capture non-linear changes, shifts and potential interactions in covariates without prespecification of such effects. As supported by the application presented on probabilistic wind direction forecasting, this can be particularly useful for modelling a highly fluctuating response, such as typically observed for wind direction, or/and in case of a large number of possible covariates. Moreover, the application shows that build-

ing ensembles of circular regression trees can even improve the forecasting performance, as the resulting forests enable modelling smooth effects and stabilize the model.

6. Computational details

The corresponding implementation of the proposed methodology for circular regression trees and forests is provided in the R package `circtree` (version 0.1.0). The package is based on the `disttree` package (version 0.2.0) which applies the main tree building functions from the `partykit` package (version 1.2.7). All three packages are available on R-Forge at `https://R-Forge.R-project.org/projects/partykit/`.

For the circular GLM that was considered as reference model the corresponding implementation is provided in the R package `circglmbayes` by Mulder and Klugkist (2017). In particular the function `circGLM` is applied to estimate the intercept and regression coefficient along with the concentration parameter.

The computation time on a standard computer is of the order of milliseconds for the circular fit (used for the persistency and the climatological model), seconds for the estimation of the linear model and the circular tree, and from a few minutes up to an hour for the estimation of the circular forest; estimation of the circular forest strongly depends on the number of covariates that are employed and, hence, varies between the two locations that were included in the application. However, for a single station and forecast step, the persistency model must be re-estimated for each time of interest, and independent climatological and linear models must be fitted for all times of interest within a single calendar year, whereas for all times of interest only a single circular tree and a single circular forest model are required. In general with respect to other applications, the computation time for the estimation of circular regression trees and forests can be reduced by various implemented settings, such as parallel estimation on multiple central processing unit cores or binning of the input data (see Section 4.2).

Acknowledgements

This project was partially funded by Austrian Research Promotion Agency grant 858537. Torsten Hothorn received funding from the Swiss National Science Foundation, grant 200021_184603. Lisa Schlosser received a doctoral scholarship granted by the University of Innsbruck.

Appendix A: Tree algorithm

This section provides a more detailed overview on the permutation-test-based CTree algorithm (Hothorn *et al.*, 2006), specifically for circular data as applied for building circular regression trees and forests presented in this study. An alternative tree building framework is provided by the MOB algorithm, which is based on M-fluctuation tests (see Zeileis *et al.* (2008) for more details).

In what follows, the testing and splitting strategy is described for the root node of the tree which contains the entire learning sample. For a complete tree model, the same procedure is applied iteratively to all resulting child nodes with the corresponding subsamples.

First, employing the von Mises distribution, a distributional model $\mathrm{vM}(y; \mu, \kappa)$ is fitted to the learning sample of circular observations $\{y_i\}_{i=1,\ldots,n}$ as explained in Section 2. In the next step, a goodness-of-fit measurement is obtained for each parameter and each observation by evaluating the score function $s(\mu, \kappa, y)$ at the estimated location and concentration parameter $\hat{\mu}$ and $\hat{\kappa}$. To detect dependences between the resulting scoring matrix

$$\begin{pmatrix} s(\hat{\mu}, \hat{\kappa}, y_1)_1, & s(\hat{\mu}, \hat{\kappa}, y_1)_2 \\ \vdots & \vdots \\ s(\hat{\mu}, \hat{\kappa}, y_n)_1, & s(\hat{\mu}, \hat{\kappa}, y_n)_2 \end{pmatrix} \tag{A.1}$$

and each possible split variable $z_l \in \{z_1, \ldots, z_m\}$ a permutation test is applied. In particular, the null hypotheses of independence of each split variable and the scores is assessed by employing the multivariate linear statistic

$$t_l = \text{vec}\left\{\sum_{i=1}^{n} v_l(z_{li})\, s(\hat{\mu}, \hat{\kappa}, y_i)\right\} \tag{A.2}$$

with $s(\hat{\mu}, \hat{\kappa}, y_i) \in \mathbb{R}^{1\times 2}$. For a numeric split variable z_l the transformation function v_l is simply the identity function $v_l(z_{li}) = z_{li}$ such that $t_l \in \mathbb{R}^2$ as the 'vec' operator converts the matrix of dimension 1×2 into a two-column vector. If z_l is a categorical variable with h categories then $v_l(z_{li}) = (\mathbf{I}(z_{li} = 1), \ldots, \mathbf{I}(z_{li} = h))$; hence, v_l returns a unit vector of dimension h where the entry 1 indicates the category of z_{li}. In this case the vec operator converts the $h \times 2$ matrix into an column vector of length $2h$ by columnwise combination such that $t_l \in \mathbb{R}^{2h}$. If there are any observations with missing values these are not included in the calculation of t_l.

To map the multivariate linear statistic t_l onto the real line a univariate test statistic c is employed; for example in a quadratic form

$$c_{\text{quad}}(t_l, \mu_l, \Sigma_l) = (t_l - \mu_l)\Sigma_l^{+}(t_l - \mu_l)^{\mathrm{T}} \tag{A.3}$$

where μ_l and Σ_l are the conditional expectation and the covariance of t_l, as derived by Strasser and Weber (1999) and used for standardization, and Σ_l^{+} is the Moore–Penrose inverse of Σ_l. As an alternative, also a maximum form ($c_{\max}$) can be considered such that the maximum of the absolute values of the standardized linear statistic is returned.

The asymptotic conditional distribution of $c(t_l, \mu_l, \Sigma_l)$ is either normal (for $c_{\max}$) or χ^2 (for c_{quad}) because the asymptotic conditional distribution of the linear statistic t_l is a multivariate normal distribution with parameters μ_l and Σ_l (Strasser and Weber, 1999). With this knowledge at hand, the corresponding p-values can be calculated and used to select the best splitting variable. A small p-value corresponding to $c(t_l, \mu_l, \Sigma_l)$ indicates a strong discrepancy from the assumption of independence between the scores and the split variable z_l. Therefore, if any of the Bonferroni-adjusted p-values is beneath the significance level selected, the partitioning variable z_{l*} with the lowest p-value is selected as split variable; otherwise no split is performed. This early stopping induced by the significance level is referred to as 'prepruning' which is often avoided for forest models by setting the significance level to 1.

To select the best split point within the already chosen split variable, again, a linear test statistic is employed. In particular, for a break point r of the variable z_{l*} leading to two subgroups $\mathcal{B}_{1r}$ and $\mathcal{B}_{2r}$ the discrepancy between score functions in the subgroups is measured by evaluating

$$t_{l*}^{qr} = \sum_{i \in \mathcal{B}_{qr}} s(\hat{\mu}, \hat{\kappa}, y_i) \tag{A.4}$$

for $q \in \{1, 2\}$. The break point that leads to the highest discrepancy is then selected as split point as defined by

$$r^* = \arg\max_r [\max_{q=1,2} \{c(t_{l*}^{qr}, \mu_{l*}^{qr}, \Sigma_{l*}^{qr})\}]. \tag{A.5}$$

Subsequently, the same testing and splitting procedure is repeated in each of the resulting subgroups until some stopping criterion has been reached. Complementing the prepruning based on the significance level of the statistical tests, as described above, further criteria such as maximal tree depth or a minimal number of observations in a node can be employed.

In addition to, or as an alternative to, prepruning, classical post-pruning can also be applied where a large tree is grown first and then pruned back based on the predictive performance (Breiman *et al.*, 1984). As pointed out by Zeileis *et al.* (2008), information criteria such as AIC (the Akaike information criterion) or BIC (the Bayes information criterion) are natural candidates for this when partitioning maximum likelihood models. However, Schlosser *et al.* (2019b) showed empirically that post-pruning only improves on prepruning in an unbiased tree algorithm when the underlying significance tests do not work well, e.g. when they have low power or when the number of observations is extremely large.

References

Breiman, L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984) *Classification and Regression Trees*. Pacific Grove. Wadsworth.

Fisher, N. I. (1993) *Statistical Analysis of Circular Data*. Cambridge: Cambridge University Press.
Fisher, N. I. and Lee, A. J. (1992) Regression models for an angular response. *Biometrics*, **48**, 665–677.
Gill, J. and Hangartner, D. (2010) Circular data in political science and how to handle it. *Polit. Anal.*, **18**, 316–336.
Gneiting, T. (2008) Editorial: Probabilistic forecasting. *J. R. Statist. Soc.* A, **171**, 319–321.
Gneiting, T. and Raftery, A. E. (2007) Strictly proper scoring rules, prediction, and estimation. *J. Am. Statist. Ass.*, **102**, 359–378.
Gould, A. L. (1969) A regression technique for angular variates. *Biometrics*, **25**, 683–700.
Grimit, E. P., Gneiting, T., Berrocal, V. J. and Johnson, N. A. (2006) The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Q. J. R. Meteorol. Soc.* C, **132**, 2925–2942.
Hara, K. and Chellappa, R. (2017) Growing regression tree forests by classification for continuous object pose estimation. *Int. J. Comput. Visn*, **122**, 292–312.
Hersbach, H. (2000) Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weath. Forecast.*, **15**, 559–570.
Hothorn, T., Hornik, K. and Zeileis, A. (2006) Unbiased recursive partitioning: a conditional inference framework. *J. Computnl Graph. Statist.*, **15**, 651–674.
Hothorn, T. and Zeileis, A. (2017) Transformation forests. *Preprint arXiv 1701.02110*. Universität Zürich, Zürich.
Jammalamadaka, S. R. and Sengupta, A. (2001) *Topics in Circular Statistics*. Singapore: World Scientific Publishing.
Johnson, R. A. and Wehrly, T. E. (1978) Some angular-linear distributions and related regression models. *J. Am. Statist. Ass.*, **73**, 602–606.
Lang, M. N., Lerch, S., Mayr, G. J., Simon, T., Stauffer, R. and Zeileis, A. (2020) Remember the past: a comparison of time-adaptive training schemes for non-homogeneous regression. *Nonlin. Process. Geophys.*, **27**, 23–34.
Lang, M. N., Mayr, G. J., Stauffer, R. and Zeileis, A. (2019) Bivariate Gaussian models for wind vectors in a distributional regression framework. *Adv. Statist. Climtol. Meteorol. Oceanog.*, **5**, no. 2, 115–132.
Ley, C. and Verdebout, T. (2017) *Modern Directional Statistics*. Boca Raton: Chapman and Hall–CRC.
Lund, U. J. (1999) Least circular distance regression for directional data. *J. Appl. Statist.*, **26**, 723–733.
Lund, U. J. (2002) Tree-based regression for a circular response. *Communs Statist. Theory Meth.*, **31**, 1549–1560.
Mardia, K. V. and Jupp, P. E. (1999) *Directional Statistics*. Chichester: Wiley.
Mardia, K. V. and Zemroch, P. J. (1975) Algorithm AS 86: The von Mises distribution function. *Appl. Statist.*, **24**, 268–272.
Mulder, K. and Klugkist, I. (2017) Bayesian estimation and hypothesis tests for a circular generalized linear model. *J. Math. Psychol.*, **80**, 4–14.
National Oceanic and Atmospheric Administration National Weather Service (2019) National Weather Service glossary. National Oceanic and Atmospheric Administration, Silver Spring. (Available from `https://w1.weather.gov/glossary/`.)
Pewsey, A., Neuhäuser, M. and Ruxton, G. D. (2013) *Circular Statistics in R*. Oxford: Oxford University Press.
Rigby, R. A. and Stasinopoulos, D. M. (2005) Generalized additive models for location, scale and shape (with discussion). *Appl. Statist.*, **54**, 507–554.
Schlosser, L., Hothorn, T., Stauffer, R. and Zeileis, A. (2019a) Distributional regression forests for probabilistic precipitation forecasting in complex terrain. *Ann. Appl. Statist.*, **13**, 1564–1589.
Schlosser, L., Hothorn, T. and Zeileis, A. (2019b) The power of unbiased recursive partitioning: a unifying view of CTree, MOB, and GUIDE. *Preprint arXiv 1906.10179*. Universität Innsbruck, Innsbruck
Simon, T., Umlauf, N., Zeileis, A., Mayr, G. J., Schulz, W. and Diendorfer, G. (2017) Spatio-temporal modelling of lightning climatologies for complex terrain. *Natrl Haz. Earth Syst. Sci.*, **17**, 305–314.
Stauffer, R., Mayr, G. J., Messner, J. W., Umlauf, N. and Zeileis, A. (2017) Spatio-temporal precipitation climatology over complex terrain using a censored additive regression model. *Int. J. Climtol.*, **37**, 3264–3275.
Strasser, H. and Weber, C. (1999) On the asymptotic theory of permutation statistics. *Math. Meth. Statist.*, **8**, 220–250.
Thorarinsdottir, T. L. and Gneiting, T. (2010) Probabilistic forecasts of wind speed: ensemble model output statistics by using heteroscedastic censored regression. *J. R. Statist. Soc.* A, **173**, 371–388.
Vogel, P., Knippertz, P., Fink, A. H., Schlüter, A. and Gneiting, T. (2018) Skill of global raw and postprocessed ensemble predictions of rainfall over Northern Tropical Africa. *Weath. Forecast.*, **33**, 369–388.
Wessel, B., Huber, M., Wohlfart, C., Marschalk, U., Kosmann, D. and Roth, A. (2018) Accuracy assessment of the Global TanDEM-X Digital Elevation Model with GPS Data. *J. Photgramm. Remote Sens.*, **139**, 171–182.
Zeileis, A., Hothorn, T. and Hornik, K. (2008) Model-based recursive partitioning. *J. Computnl Graph. Statist.*, **17**, 492–514.

Article XI

Stauffer R., and Zeileis A. (2024): *colorspace: A Python Toolbox for Manipulating and Assessing Colors and Palettes*. Journal of Open Source Software, 9(102), 7120, doi:10.21105/joss.07120.

Recent open-source and open-review software journal, not yet listed in JCR.

The Journal of Open Source Software

colorspace: A Python Toolbox for Manipulating and Assessing Colors and Palettes

Reto Stauffer [1,2] **and Achim Zeileis** [1]

1 Department of Statistics, Universität Innsbruck, Austria **2** Digital Science Center, Universität Innsbruck, Austria

DOI: 10.21105/joss.07120

Software

- Review
- Repository
- Archive

Editor: Julia Romanowska

Reviewers:

- @hollowscene
- @dmreagan

Submitted: 29 July 2024
Published: 29 October 2024

Summary

The Python *colorspace* package provides a toolbox for mapping between different color spaces, which can then be used to generate a wide range of perceptually-based color palettes for qualitative or quantitative (sequential or diverging) information. These palettes (as well as any other sets of colors) can be visualized, assessed, and manipulated in various ways, e.g., by color swatches, emulating the effects of color vision deficiencies, or depicting the perceptual properties. Finally, *colorspace* integrates seamlessly with standard Python graphics packages like *matplotlib*, *seaborn*, and *plotly*, making it a valuable resource for both developers and practitioners to customize, assess, and implement color palettes in their data visualization workflows.

Statement of need

Color is an integral element of visualizations and graphics and is essential for communicating (scientific) information. However, colors need to be chosen carefully so that they support the information displayed for all viewers (see e.g., Tufte, 1990; Ware, 2004; Wilke, 2019). Therefore, suitable color palettes have been proposed in the literature (e.g., Brewer, 1999; Crameri et al., 2020; Ihaka, 2003) and many software packages transitioned to better color defaults over the last decade. A prominent example from the Python community is *matplotlib* 2.0 (Hunter et al., 2017), which replaced the classic "jet" palette (a variation of the infamous "rainbow") by the perceptually-based "viridis" palette. Hence a wide range of useful palettes for different purposes is provided in a number of Python packages today, including *cmcramery* (Rollo, 2024), *colormap* (Cokelaer, 2024), *colormaps* (Patel, 2024), *matplotlib* (Hunter, 2007), *palettable* (Davis, 2023), and *seaborn* (Waskom, 2021).

However, colors are provided as a fixed set in most graphics packages. While this makes it easy to use them in different applications, it is usually not easy to modify the perceptual properties or to set up new palettes following the same principles. The *colorspace* package addresses this by supporting color descriptions using different color spaces (hence the package name), including some that are based on human color perception. One notable example is the Hue-Chroma-Luminance (HCL) model, which represents colors by coordinates on three perceptually-based axes: hue (type of color), chroma (colorfulness), and luminance (brightness). Selecting colors along paths along these axes allows for intuitive construction of palettes that closely match many of the palettes provided in the packages listed above.

In addition to functions and interactive apps for HCL-based colors, the *colorspace* package also offers functions and classes for handling, transforming, and visualizing color palettes (from any source). In particular, this includes the simulation of color vision deficiencies (Machado et al., 2009) but also contrast ratios, desaturation, lightening/darkening, etc.

The *colorspace* Python package was inspired by the eponymous R package (Zeileis et al., 2020). It comes with extensive documentation at https://retostauffer.github.io/python-colorspace/, including many practical examples. The package complements existing graphics packages in Python both for casual users and data visualization experts. Selected highlights are presented in the following, motivating its usefulness for various kinds of graphics in different fields of application and research.

Key functionality

HCL-based color palettes

The key functions and classes for constructing color palettes using hue-chroma-luminance paths (and then mapping these to hex codes) are:

- `qualitative_hcl`: For qualitative or unordered categorical information, where every color should receive a similar perceptual weight.
- `sequential_hcl`: For ordered/numeric information from high to low (or vice versa).
- `diverging_hcl`: For ordered/numeric information around a central neutral value, where colors diverge from neutral to two extremes.

These functions provide a range of named palettes inspired by well-established packages but actually implemented using HCL paths. Additionally, the HCL parameters can be modified or new palettes can be created from scratch.

As an example, Figure 1 depicts color swatches for four viridis variations. The first, `pal1`, sets up the palette from its name. It is identical to the second, `pal2`, which employes the HCL specification directly: the hue ranges from purple (300) to yellow (75), colorfulness (chroma) increases from 40 to 95, and luminance (brightness) from dark (15) to light (90). The `power` parameter chooses a linear change in chroma and a slightly nonlinear path for luminance.

In `pal3` and `pal4`, the most HCL properties are kept the same but some are modified: `pal3` uses a triangular chroma path from 40 via 90 to 20, yielding muted colors at the end of the palette. `pal4` just changes the starting hue for the palette to green (200) instead of purple. All four palettes are visualized by the `swatchplot` function from the package.

Figure 1: Swatches of four HCL-based sequential palettes: `pal1` is the predefined HCL-based viridis palette, `pal2` is identical to `pal2` but created "by hand" and `pal3` and `pal4` are modified versions with a triangular chroma paths and reduced hue range, respectively.

The objects returned by the palette functions provide a series of methods, e.g., `pal1.settings` for displaying the HCL parameters, `pal1(3)` for obtaining a number of hex colors, or `pal1.cmap()` for setting up a *matplotlib* color map, among others.

```
from colorspace import palette, sequential_hcl, swatchplot
pal1 = sequential_hcl(palette = "viridis")
pal2 = sequential_hcl(h = [300, 75], c = [40, 95], l = [15, 90],
                      power = [1., 1.1])
pal3 = sequential_hcl(palette = "viridis", cmax = 90,  c2 = 20)
```

```
pal4 = sequential_hcl(palette = "viridis", h1 = 200)
swatchplot({"Viridis (and altered versions of it)": [
               palette(pal1(7), "By name"),
               palette(pal2(7), "By hand"),
               palette(pal3(7), "With triangular chroma"),
               palette(pal4(7), "With smaller hue range")
           ]}, figsize = (8, 1.75));
```

An overview of the named HCL-based palettes in *colorspace* is depicted in Figure 2.

```
from colorspace import hcl_palettes
hcl_palettes(plot = True, figsize = (20, 15))
```

Figure 2: Overview of the predefined (fully customizable) HCL color palettes.

Palette visualization and assessment

To better understand the properties of palette `pal4`, defined above, Figure 3 shows its HCL spectrum (left) with separate lines for the hue, chroma, and luminance coordinates and the corresponding path through the three-dimensional HCL space (right) where hue co-varies along with chroma and luminance.

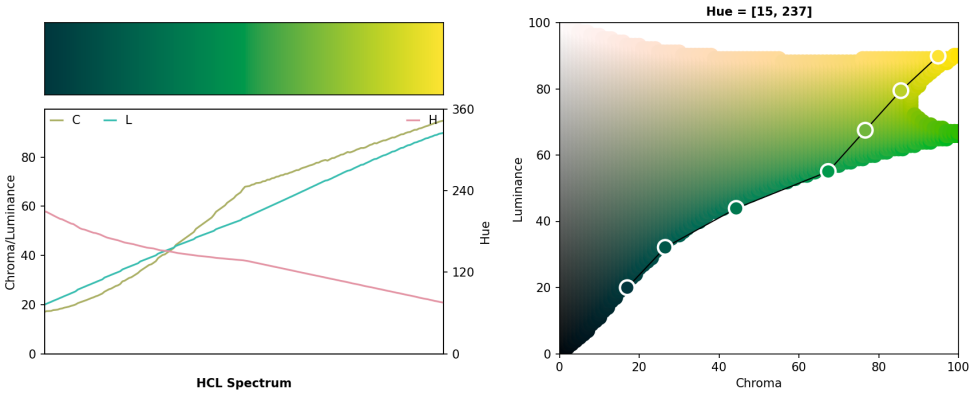

Figure 3: Hue-chroma-luminance spectrum plot (left) and corresponding path in the chroma-luminance coordinate system (where hue changes with luminance) for the custom sequential palette `pal4`.

The spectrum in the first panel shows how the hue (right axis) changes from about 200 (green) to 75 (yellow), while chroma and luminance (left axis) increase from about 20 to 95. Note that the kink in the chroma curve for the greenish colors occurs because such dark greens cannot have higher chromas when represented through RGB-based hex codes. The same is visible in the second panel where the path moves along the outer edge of the HCL space.

```
pal4.specplot(figsize = (5, 5));
pal4.hclplot(n = 7, figsize = (5, 5));
```

Color vision deficiency

Another important assessment of a color palette is how well it works for viewers with color vision deficiencies. This is exemplified in Figure 4, which depicts a demo plot (heatmap) under "normal" vision (left), deuteranomaly (colloquially known as "red-green color blindness", center), and desaturated (gray scale, right). The palette in the top row is the traditional fully-saturated RGB rainbow, deliberately selected here as a palette with poor perceptual properties. It is contrasted with a perceptually-based sequential blue-yellow HCL palette in the bottom row.

The sequential HCL palette is monotonic in luminance so that it is easy to distinguish high-density and low-density regions under deuteranomaly and desaturation. However, the rainbow is non-monotonic in luminance and parts of the red-green contrasts collapse under deuteranomaly, making it much harder to interpret correctly.

```
from colorspace import rainbow, sequential_hcl
col1 = rainbow(end = 2/3, rev = True)(7)
col2 = sequential_hcl("Blue-Yellow", rev = True)(7)

from colorspace import demoplot, deutan, desaturate
import matplotlib.pyplot as plt
fig, ax = plt.subplots(2, 3, figsize = (9, 4))
demoplot(col1, "Heatmap", ax = ax[0,0], ylabel = "Rainbow", title = "Original")
demoplot(col2, "Heatmap", ax = ax[1,0], ylabel = "HCL (Blue-Yellow)")
demoplot(deutan(col1), "Heatmap", ax = ax[0,1], title = "Deuteranope")
demoplot(deutan(col2), "Heatmap", ax = ax[1,1])
demoplot(desaturate(col1), "Heatmap", ax = ax[0,2], title = "Desaturated")
demoplot(desaturate(col2), "Heatmap", ax = ax[1,2])
plt.show()
```

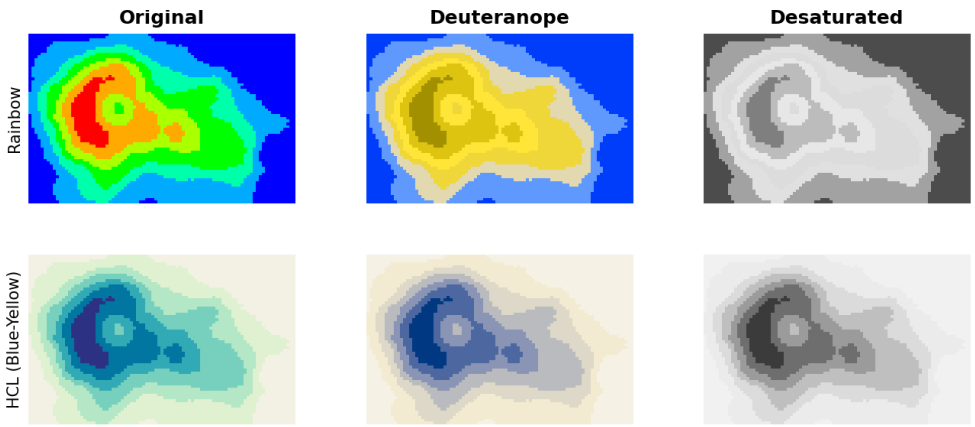

Figure 4: Example of color vision deficiency emulation and color manipulation using a heatmap. Top/bottom: RGB rainbow based palette and HCL based sequential palette. Left to right: Original colors, deuteranope color vision, and desaturated representation.

Integration with Python graphics packages

To illustrate that *colorspace* can be easily combined with different graphics workflows in Python, Figure 5 shows a heatmap (two-dimensional histogram) from *matplotlib* and multi-group density from *seaborn*. The code below employs an example data set from the package (using *pandas*) with daily maximum and minimum temperature. For *matplotlib* the colormap (`.cmap()`; `LinearSegmentedColormap`) is extracted from the adapted viridis palette `pal3` defined above. For *seaborn* the hex codes from a custom qualitative palette are extracted via `.colors(4)`.

```
from colorspace import dataset, qualitative_hcl
import matplotlib.pyplot as plt
import seaborn as sns

df = dataset("HarzTraffic")

fig = plt.hist2d(df.tempmin, df.tempmax, bins = 20,
                 cmap = pal3.cmap().reversed())
plt.title("Joint density daily min/max temperature")
plt.xlabel("minimum temperature [deg C]")
plt.ylabel("maximum temperature [deg C]")
plt.show()

pal = qualitative_hcl("Dark 3", h1 = -180, h2 = 100)
g = sns.displot(data = df, x = "tempmax", hue = "season", fill = "season",
                kind = "kde", rug = True, height = 4, aspect = 1,
                palette = pal.colors(4))
g.set_axis_labels("temperature [deg C]")
g.set(title = "Distribution of daily maximum temperature given season")
plt.show()
```

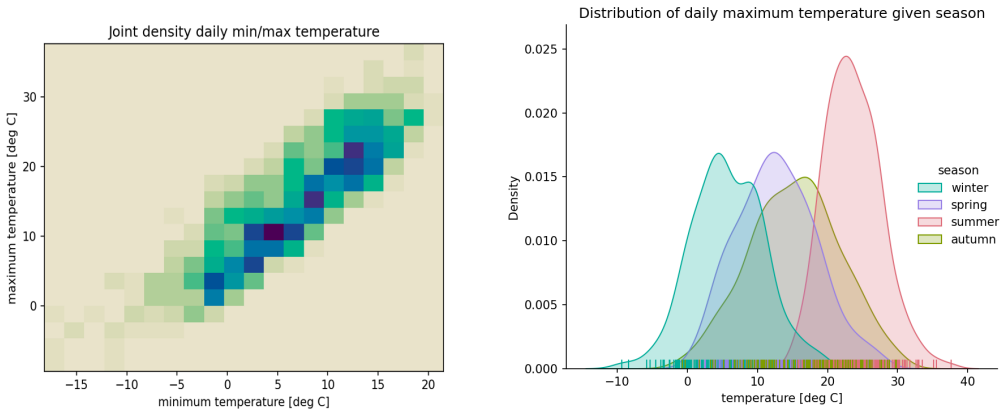

Figure 5: Example of a `matplotlib` heatmap and a `seaborn` density using custom HCL-based colors.

Dependencies and availability

The *colorspace* package is available from PyPI at https://pypi.org/project/colorspace. It is designed to be lightweight, requiring only *numpy* (Harris et al., 2020) for the core functionality. Only a few features rely on *matplotlib*, *imageio* (Klein et al., 2024), and *pandas* (The Pandas Development Team, 2024). More information and an interactive interface can be found on https://hclwizard.org/. Package development is hosted on GitHub at https://github.com/retostauffer/python-colorspace. Bug reports, code contributions, and feature requests are warmly welcome.

References

Brewer, C. A. (1999). Color use guidelines for data representation. *Proceedings of the Section on Statistical Graphics, American Statistical Association*, 55–60.

Cokelaer, T. (2024). *Colormap* (Version v1.1.0). Python Package Index (PyPI). https://pypi.org/project/colormap/

Crameri, F., Shephard, G. E., & Heron, P. J. (2020). The misuse of colour in science communication. *Nature Communications*, *11*(5444), 1–10. https://doi.org/10.1038/s41467-020-19160-7

Davis, M. (2023). *palettable: Color palettes for Python* (Version v3.3.3). Python Package Index (PyPI). https://pypi.org/project/palettable/

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/mcse.2007.55

Hunter, J. D., Dale, D., Firing, E., Droettboom, M., & the Matplotlib Development Team. (2017). *What's new in matplotlib 2.0, changes to the default style*. https://matplotlib.org/stable/users/prev_whats_new/dflt_style_changes.html

Ihaka, R. (2003). Colour for presentation graphics. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd international workshop on distributed statistical computing, vienna, austria*. https://www.R-project.org/conferences/DSC-2003/Proceedings/Ihaka.pdf

Klein, A., Wallkötter, S., Silvester, S., Rynes, A., actions-user, Müller, P., Nunez-Iglesias, J., Harfouche, M., Schrangl, L., Dennis, Lee, A., Pandede, McCormick, M., OrganicIrradiation, Rai, A., Ladegaard, A., van Kemenade, H., Smith, T. D., Vaillant, G., … Singleton, J. (2024). *Imageio/imageio* (Version v2.34.2). Zenodo. https://doi.org/10.5281/zenodo.12514964

Machado, G. M., Oliviera, M. M., & Fernandes, L. A. F. (2009). A physiologically-based model for simulation of color vision deficiency. *IEEE Transactions on Visualization and Computer Graphics*, *15*(6), 1291–1298. https://doi.org/10.1109/tvcg.2009.113

Patel, P. (2024). *Colormaps* (Version v0.4.2). Python Package Index (PyPI). https://pypi.org/project/colormaps/

Rollo, C. (2024). *cmcrameri: Python wrapper around Fabio Crameri's perceptually uniform colormaps* (Version v1.9). Python Package Index (PyPI). https://pypi.org/project/cmcrameri/

The Pandas Development Team. (2024). *pandas-Dev/Pandas: pandas* (Version v2.2.2). Zenodo. https://doi.org/10.5281/zenodo.10957263

Tufte, E. (1990). *Envisioning information*. Graphics Press.

Ware, C. (2004). Color. In *Information visualization: Perception for design* (pp. 103–149). Morgan Kaufmann Publishers Inc.

Waskom, M. L. (2021). seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Wilke, C. O. (2019). *Fundamentals of data visualization*. O'Reilly Media. ISBN: 1492031089

Zeileis, A., Fisher, J. C., Hornik, K., Ihaka, R., McWhite, C., Murrell, P., Stauffer, R., & Wilke, C. O. (2020). colorspace: A toolbox for manipulating and assessing colors and palettes. *Journal of Statistical Software*, *96*(1), 1–49. https://doi.org/10.18637/jss.v096.i01

Article XII

Stauffer R., Zeileis A., and Mayr G.J. (2024). *Long-Term Foehn Reconstruction Combining Unsupervised and Supervised Learning*. International Journal of Climatology, *forthcoming*, doi:10.1002/joc.8673.

JCR ranking: **Category 1** in *Meteorology & Atmospheric Sciences*.

International Journal of Climatology

International Journal of Climatology RMetS

RESEARCH ARTICLE OPEN ACCESS

Long-Term Foehn Reconstruction Combining Unsupervised and Supervised Learning

Reto Stauffer[1] | Achim Zeileis[2] | Georg J. Mayr[3]

[1]Faculty of Economics and Statistics & Digital Science Center, Universität Innsbruck, Innsbruck, Austria | [2]Faculty of Economics and Statistics, Universität Innsbruck, Innsbruck, Austria | [3]Department of Atmospheric and Cryospheric Sciences, Universität Innsbruck, Innsbruck, Austria

Correspondence: Reto Stauffer (reto.stauffer@uibk.ac.at)

Received: 3 June 2024 | **Revised:** 29 August 2024 | **Accepted:** 20 October 2024

Keywords: climate | foehn | mixture model | reconstruction | supervised | trend | unsupervised

ABSTRACT

Foehn winds, characterised by abrupt temperature increases and wind speed changes, significantly impact regions on the leeward side of mountain ranges, e.g., by spreading wildfires. Understanding how foehn occurrences change under climate change is crucial. As foehn is a meteorological phenomenon, its prevalence has to be inferred from meteorological measurements employing suitable classification schemes. Hence, this approach is typically limited to specific periods for which the necessary data are available. We present a novel approach for reconstructing historical foehn occurrences using a combination of unsupervised and supervised probabilistic statistical learning methods. We utilise in situ measurements (available for recent decades) to train an unsupervised learner (finite mixture model) for automatic foehn classification. These labelled data are then linked to reanalysis data (covering longer periods) using a supervised learner (lasso or boosting). This allows us to reconstruct past foehn probabilities based solely on reanalysis data. Applying this method to ERA5 reanalysis data for six stations across Switzerland and Austria achieves accurate hourly reconstructions of north and south foehn occurrence, respectively, dating back to 1940. This paves the way for investigating how seasonal foehn patterns have evolved over the past 83 years, providing valuable insights into climate change impacts on these critical wind events.

1 | Introduction

Foehn winds are downslope winds on the leeward side of mountains and can be found all around the world in areas with pronounced topographical features that impede the airflow, such as the European Alps, the Southern Alps in New Zealand, mountain ranges along the Mediterranean Sea or the Rocky Mountains. Depending on the region, these winds are given specific names such as Santa Anna winds (Southern California; Sergius, Ellis, and Ogden 1962; Rolinski, Capps, and Zhuang 2019), Chinook (Rocky Mountains; Armi and Mayr 2015), Bora (Croatia; Grisogono and BelušIć 2009), Zonda (Andes, Argentina; Norte 2015), Raco (Chile; Muñoz and Armi 2024), Jintsu-Oroshi and Inami-Kaze (Japan; Kusaka et al. 2021; Koyanagi and Kusaka 2020), Halny (Tatra Mountains, Poland; Śliwińska and Ciaranek 2015; Grajek and Bednorz 2024) or Foehn (Central Europe and New Zealand; McGowan and Sturman 1996; Richner and Hächler 2013; McClung and Mass 2020). For historical reasons, 'foehn' has become a synonym for this type of terrain-induced wind phenomena.

Often, foehn is characterised by a sharp increase in wind speed and sudden changes in temperature and relative humidity, which can have a strong influence on the local climate and the people living in the affected areas. While foehn is often associated with a mild (and typically dry) climate, strong foehn events can also cause extensive damage to vegetation and man-made structures. In some areas, it is not uncommon for strong foehn gusts to overturn trucks or vans, or for airports and harbours to be closed due to unsafe conditions. In addition, the strong and

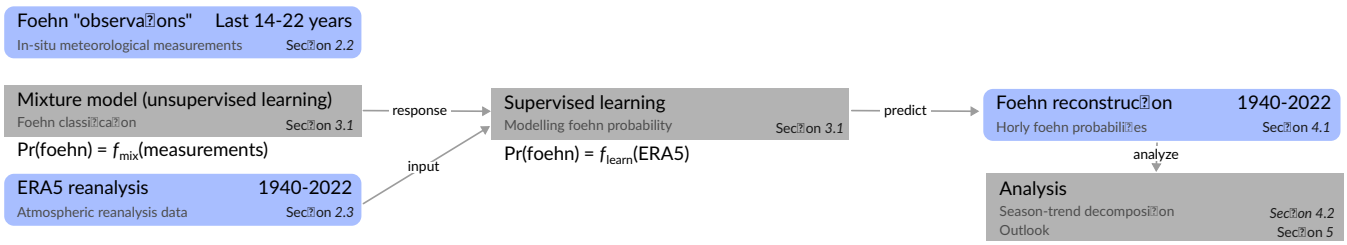

FIGURE 1 | Flow chart for the new combined foehn reconstruction algorithm. [Colour figure can be viewed at wileyonlinelibrary.com]

dry winds can kindle and spread domestic fires and wildfires (Schoennagel, Veblen, and Romme 2004; Reinhard, Rebetez, and Schlaepfer 2005; Zumbrunnen et al. 2009) or affect the development of Antarctic ice shelves (e.g., Cape et al. 2015; Elvidge et al. 2020).

While the conceptual model of foehn is well established (Armi and Mayr 2007, 2011; Mayr and Armi 2008; Richner and Hächler 2013), its prevalence must be derived from its physical quantities such as wind, temperature and relative humidity. During the last decades, several (semi)-automatic methods have been developed that allow the differentiation of 'foehn' and 'no-foehn' events based on in situ measurements from automated weather stations (AWSs). Among the frequently used algorithms are Widmer's föhn index (Widmer 1966; Courvoisier and Gutermann 1971) based on Fisher's linear discriminant analysis to distinguish between two or more distinct classes and enhanced versions of it (e.g., Jansing et al. 2022). Other studies use decision-based or tree-based methods to classify foehn events (e.g., Dürr 2008; Speirs et al. 2013; Cape et al. 2015; Turton et al. 2018; Datta et al. 2019; Elvidge et al. 2020; Laffin et al. 2021; Francis et al. 2023), all of which are deterministic methods, where the thresholds have often been selected manually. To overcome these limitations, Plavcan, Mayr, and Zeileis (2014) proposed a method based on finite mixture models for automatic and fully parametric probabilistic foehn classification. To perform the classification, all methods require AWS measurements with high temporal resolution (ideally sub-hourly), which are typically only available for recent decades. While this allows for the classification and analysis of foehn when the AWS provides sufficient data, it does not offer information on foehn prior to the installation of the AWS, nor during outages after the AWS has been decommissioned.

Additional information on the atmospheric conditions from numerical (re-)analysis models is an excellent source to complement the in situ measurements. Reanalysis data sets are typically produced by physically based numerical weather prediction (NWP) models and sophisticated data assimilation schemes that use all available observations to estimate the 'best known' atmospheric state. An example is the global reanalysis data set ERA5 (Hersbach et al. 2023a, 2023b) from the European Centre for Medium-Range Weather Forecasts (ECMWF). ERA5 provides global hourly four-dimensional atmospheric conditions with a horizontal resolution of $_{0.25°\times0.25°}$ ($\approx 30 \times 30$ km at the equator) back to 1940. However, this comes with its own challenges. Due to technical and computational limitations, ERA5 and other NWPs with a similar (or coarser) resolution can only approximate the real world and cannot resolve small-scale atmospheric processes and topographic features, which are important for small-scale phenomena such as foehn.

One way to overcome these limitations is to combine AWS measurements and reanalysis data using statistical or machine learning techniques. A classification of foehn based on AWS data serves as the response (target/outcome/labels) for a supervised model that uses reanalysis data as explanatory variables (inputs/covariates). Once the relationship between the two sets of data has been learned, the models can be used to predict the expected state ('foehn' or 'no foehn') for periods for which reanalysis data is available, but AWS measurements are not. This is also known as (statistical) downscaling or post-processing and has been used for foehn modelling in some variations. For nowcasting foehn at a station in Switzerland (Altdorf), Sprenger et al. (2017) use data from a local 7 km analysis data set (COSMO-7) for their adaptive boosting algorithm (AdaBoost). Laffin et al. (2021) use both ERA5 and the Regional Atmospheric Climate Model 2 (RACMO2; Wessen and Laffin 2022) combined with tree-based gradient boosting (XGBoost; Chen and Guestrin 2016) to predict foehn on the Antarctic Peninsula. XGBoost is also used by Mony, Jansing, and Sprenger (2021) to investigate future changes to foehn frequency in Switzerland.

This study proposes a novel probabilistic approach that combines unsupervised and supervised machine learning methods to bridge the gap between in situ automatic weather station (AWS) measurements and ERA5 reanalysis data to diagnose foehn. This combined approach allows us to reconstruct long-term, high-resolution foehn conditions over several decades. ERA5 data enables us to reconstruct the probability of foehn occurrence with hourly temporal resolution dating back to 1940, long before AWS was installed.

Our approach also allows for the identification of potential long-term changes in foehn occurrence in the European Alps from this high-resolution reconstruction. The validity of this combined approach is demonstrated by applying it to six stations located both north and south of the main Alpine ridge to show the method's effectiveness for both north and south foehn wind situations.

Figure 1 shows a schematic representation of the proposed approach. For all six locations, data are available from an AWS at the target location as well as from a nearby mountain station for the last 14–22 years (Section 2.1) on a 10 min temporal interval. A Gaussian mixture model (unsupervised learning; Section 3.1) is used for foehn classification. The result is then aggregated into binary time series ('foehn'/'no foehn') with an hourly temporal resolution to match the resolution of the ERA5 data used in the next

step. After combining the different data sources, supervised learning (Section 3.2) is used to find the relationship between a variety of interpolated and derived variables from ERA5 (Section 2.2) and the classified events. Once these statistical models have been estimated, foehn can be reconstructed (Section 3.3) for the whole period from 1940 to 2022, allowing the investigation of possible trends and/or seasonal changes over the past decades (Section 3.4).

2 | Data

Section 2.1 describes the measurement data utilised for foehn classification along with the study area and the target stations. Section 2.2 explains the reanalysis data set and its preprocessing for the supervised learning method, along with the reconstruction process.

2.1 | In Situ Measurements

This study utilises data from six AWSs situated across Switzerland and the western part of Austria, all positioned at the bottom of valleys known to be affected by foehn winds. Four of these stations are located north of the main Alpine ridge, while two are located in the canton of Ticino (Switzerland) south of the main Alpine ridge. Whilst the stations north of the Alps are prone to south foehn, the stations south of the Alps are known for the presence of north foehn.

An additional AWS upstream near the crest of the main Alpine range improves the accuracy of the foehn classification (Plavcan, Mayr, and Zeileis 2014). The stations Innsbruck and Ellbögen utilise data from Sattelberg, and the remaining four stations in Switzerland use observations from station Gütsch.

All stations provide data on mean wind speed, wind direction, air temperature and relative humidity at a 10-min temporal resolution. Table 1 displays the locations and data availability of these stations, while Figure 2 depicts a map illustrating their geographical position and the surrounding topography.

2.2 | ERA5 Reanalysis

This study makes use of ERA5 reanalysis data, which is publicly accessible via the Copernicus climate data store (Hersbach et al. 2023a, 2023b). ERA5 offers four-dimensional gridded data with an hourly temporal resolution (starting from 1940) on a spatial grid of $0.25^\circ \times 0.25^\circ$ ($\sim 28 \times 20$ km for Central Europe).

The data from 90 different fields (30 single-level fields and 60 pressure-level fields; see Tables S1 and S2 in Appendix S1) are bilinearly interpolated to the geographical location of the six stations of interest. Based on the 90 interpolated values, a series of derived variables is calculated, including vertical temperature gradients and level thickness, resulting in a total of 155 variables. These 155 variables are referred to as the 'direct' variable set as they solely rely on information retrieved directly from the geographical location of the corresponding target station.

Since foehn results from atmospheric conditions on a scale larger than the station scale (e.g., McGowan and Sturman 1996; Mayr and Armi 2010; Richner and Hächler 2013; Armi and Mayr 2015; Kusaka et al. 2021; Stoev, Post, and Guerova 2022), relying solely on data at the target location might be insufficient. Therefore, additional information from the surrounding area is incorporated by extracting data from ERA5 at a series of neighbouring points arranged in a 'star' formation around the target location. Figure 2 shows the target locations (`C`; center) and their neighbouring points used. These neighbouring points are positioned geographically relative to the target station upstream (`U`) and downstream (`D`) of the main foehn wind direction as well as to the right (`R`) and left (`L`) of it.

While the interpolated information from the target location itself (`C`) is always used as possible covariate for the statistical models, the values interpolated at the neighbouring points are not directly employed but are instead used for the calculation of the derived/augmented variables. In combination with the 'direct' variables, a list of additional derived variables is calculated such as spatio-temporal temperature and pressure differences. For example, spatial differences in surface pressure are

TABLE 1 | Station type, location and data availability; begin and end date plus percent available within the period.

	Type	Location	Data availability
△ Gütsch (Andermatt)[a]	Crest	46.653 N/8.616E 2286 m	2005-01-01–2023-12-31 (95.3%)
Altdorf[a]	South	46.890 N/8.620E 438 m	2005-01-01–2022-12-31 (78.1%)
Montana[a]	South	46.290 N/7.460E 1423 m	2005-01-01–2022-12-31 (77.1%)
Comprovasco[a]	North	46.460 N/8.935E 576 m	2005-01-01–2022-12-31 (85.8%)
Lugano[a]	North	46.004 N/8.960E 205 m	2005-01-01–2022-12-31 (90.2%)
△ Sattelberg[b]	Crest	47.011 N/11.479E 2107 m	2006-01-01–2022-12-31 (75.0%)
Ellbögen[b]	South	47.200 N/11.430E 1080 m	2006-01-01–2022-12-31 (92.0%)
(Universität) Innsbruck[c]	South	47.260 N/11.385E 578 m	2009-06-21–2022-12-30 (99.1%)

Note: Observations provided by the Swiss national weather service. Four stations are used to model south foehn, two to model north foehn and two serving information at the mountain crest (△; cf. Type).
[a]MeteoSwiss, the University of Innsbruck.
[b]The Austrian national weather service.
[c]GeoSphere Austria.

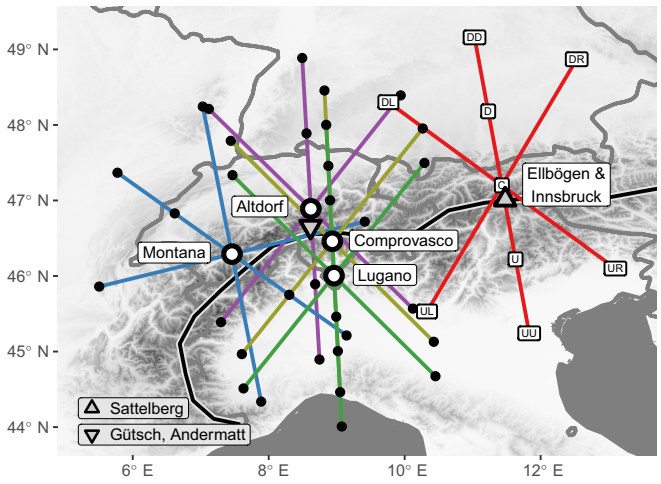

FIGURE 2 | Study area with the location of the six target stations (circles/C) and two crest stations (triangles); innsbruck and Ellbögen are shown combined due to their close proximity (< 8 km). The solid black lines represent the main Alpine ridge. On top, the neighbourhood 'star' used for interpolation is shown, exemplarily labelled on the most eastern station. [Colour figure can be viewed at wileyonlinelibrary.com]

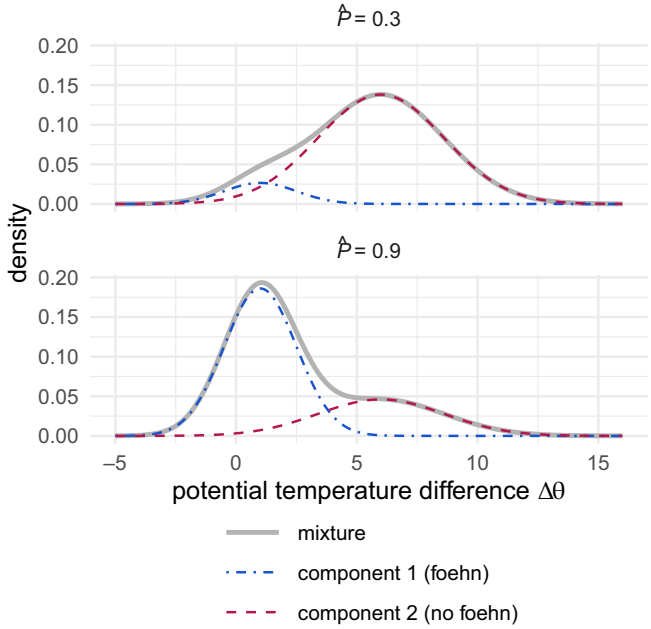

FIGURE 3 | Illustration of the two-component Gaussian mixture model used for classification. The potential temperature difference ($\Delta\theta$) is used as the main variable to separate component 1 (foehn; dot-dashed, blue) with $\mu_1 = 1.0, \sigma_1 = 1.5$ and component 2 (no foehn; dashed, red) with $\mu_2 = 6.0, \sigma_2 = 2.6$; the resulting mixture (solid, grey) depends on the posterior probability $\hat{p}$ returned by the concomitant model including relative humidity (rh) and mean wind speed (ff) at the valley station. Two examples are shown with $\hat{p} = 0.3$ and $\hat{p} = 0.9$, respectively. [Colour figure can be viewed at wileyonlinelibrary.com]

calculated between neighbouring points (e.g., *C–D*, *U–C*, *UU–DD*, *UL–DR*, *UR–DL*; see Figure 2) as well as temporal changes over the 3–6 h at these neighbouring points. Taking spatial differences of these temporal changes yields the spatio-temporal information. Finally, the first and second-order harmonics of the day of the year are included to capture seasonal variation. In total, this yields 497 variables: Four harmonics, 155 direct variables, 136 spatial variables, 120 temporal variables and 82 spatio-temporal variables. This expanded set is referred to as the 'full' variable set. Further details regarding the construction of the neighbouring 'star' can be found in Appendix S1, for a comprehensive list of all variables see the materials at https://doi.org/10.48323/gdkr5-7tt45.

3 | Methodology

Section 3.1 introduces the unsupervised learning model used for foehn classification, followed by the supervised learning models in Section 3.2. The results from Section 3.2 are then used to reconstruct hourly foehn occurrence over the past decades (Section 3.3), which is analysed employing season-trend decomposition in Section 3.4.

3.1 | Unsupervised Learning: Mixture Model for Foehn Classification

As direct foehn measurement do not exist, the data from the AWSs are currently unlabelled. Therefore, a mixture model is employed to distinguish between the 'foehn' and 'no foehn' events:

$$\Pr_{\text{obs}}(\text{foehn}) = f_{\text{mix}}(\text{measurements}) \quad (1)$$

$\Pr_{\text{obs}}(\text{foehn})$ denotes the posterior probability for a 'foehn' observation at a specific time and station, which is modelled as a function ($f_{\text{mix}}()$) of the 10-min in situ measurements (Section 2.1). We employ a two-component Gaussian mixture model (Grün and Leisch 2008) with concomitants, closely following the method proposed by Plavcan, Mayr, and Zeileis (2014) implemented in the *R* package foehnix (Stauffer 2023).

The prerequisite condition for an observation to be used for classification is that the wind direction at the location of interest falls within the prevailing foehn direction at the target location, and that wind from a specific direction is also prevalent at the corresponding crest station at the same time (details in Table S3). Only the periods matching this precondition are used for estimating the Gaussian mixture model, while $\Pr_{\text{obs}}(\text{foehn}) = 0$ is set for all remaining observations.

The underlying concept involves that two unobservable Gaussian components (or clusters) exist, one describing 'foehn' conditions and the other 'no foehn' conditions. To distinguish between these two components, a main covariate is required. In this study, the potential temperature difference ($\Delta\theta$) between the valley station (t_{valley}) and the crest station (t_{crest}) is computed with the dry-adiabatic lapse rate ($0.01\,\frac{K}{m}$), yielding $\Delta\theta = t_{\text{crest}} + 0.01\Delta h - t_{\text{valley}}$, where Δh is the height difference between the two stations (cf. Table 1 and S3). This simplification eliminates the need for air pressure measurements at both sites, which is otherwise required when using the definition of potential temperature. During foehn events, the air descends on the leeward side of the mountains, with a potential temperature difference close to zero in the absence of significant diabatic processes, such as turbulent mixing or air entrainment from tributary valleys.

However, the potential temperature difference alone might not be sufficient to adequately separate the two states. Therefore, an additional concomitant model is employed to weigh the two components conditional on additional covariates. In this study, binary logistic regression is used for the concomitant model, employing relative humidity and mean wind speed as additional covariates. Models with this specification have been shown to work well for stations in the Alpine region (e.g., Plavcan, Mayr, and Zeileis 2014; Plavcan and Mayr 2015). Figure 3 provides an illustration of this model, depicting the use of $\Delta\theta$ to separate the two components ('foehn'/'no foehn') and the effect of the concomitant model on the joint density. More details can be found in Appendix S2.

Once estimated, the mixture model provides the posterior probability $\mathrm{Pr}_{\mathrm{obs}}(\mathrm{foehn})$ for each 10 min interval where the required in situ measurements are available (Equation 1). As these foehn 'observations' need to be combined with ERA5 in the next step, the results are upscaled to an hourly temporal resolution ($\mathrm{Pr}_{1\mathrm{h}}$). Therefore, the following basic assumption is applied: Each hour is considered a 'foehn' event if at least half of the 10 min posterior probabilities within that hour (e.g., 0010–0100 UTC for 0100 UTC) are ≥ 0.5, otherwise the hour is considered a 'no foehn' event. If fewer than 4 out of 6 individual 10 min probabilities are available, the hour is excluded. This is similar to Gutermann et al. (2012) and Mony, Jansing, and Sprenger (2021) who employ a 'four out of six rule'.

$$\mathrm{Pr}_{1\mathrm{h}} = \begin{cases} \text{missing} & \text{if } \sum \mathrm{Pr}_{\mathrm{obs}} \in [0-1] < 4 \\ \text{foehn} & \text{if } \frac{1}{N} \sum \left(\mathrm{Pr}_{\mathrm{obs}} \geq 0.5\right) \geq 0.5 \\ \text{no foehn} & \text{else} \end{cases} \tag{2}$$

3.2 | Supervised Learning: Modelling Foehn Probability

The outcome of the previous section is a binary time series, with each hour labelled as either a 'foehn' or 'no foehn' observation (Equation 2). This serves as the response variable in the supervised machine learning model, using the ERA5 data (Section 2.2) as input variables. The goal is to capture the relationship between the two data sets with the model of the form

$$\mathrm{Pr}_{1\mathrm{h}}(\mathrm{foehn}) = f_{\mathrm{learn}}(\mathrm{ERA5}), \tag{3}$$

where the probability $\mathrm{Pr}_{1\mathrm{h}}$ of the binary response is modelled as a function $f_{\mathrm{learn}}()$ of the available information extracted from ERA5. $f_{\mathrm{learn}}()$ can be any learner suitable for a binary response such as logistic regression, decision trees, random forests, or neural networks, to mention a few. For this study, three different learners/models are employed (details in Appendix S3):

lasso Logistic regression with lasso (L1) regularisation (Friedman, Hastie, and Tibshirani 2010; Tay, Narasimhan, and Hastie 2023).

stabsel Logistic regression with lasso-based stability selection (Meinshausen and Bühlmann 2010).

xgboost Extreme gradient boosting (Chen and Guestrin 2016).

In order to investigate the possible benefits of incorporating large-scale information from the stations' neighbourhood, two variations of each learner are considered: One utilising the 'full' set of 497 variables and one only using the 'direct' set of 155 variables (Section 2.2).

To account for location and time of day, separate models are estimated for each of the six stations for each hour of the day (0000 UTC, 0100 UTC, ..., 2300 UTC), resulting in a total of 864 models. Depending on the station, the training data for these models include 10–18 years of data (see Table 1). In addition, a six-fold cross-validation (CV) is performed using a fixed period of 12 years (2011–2022) where, in each fold, two consecutive years are left out as test data. Ellbögen and Innsbruck are missing one fold (with test data 2013–2014) where no measurements from the crest station are available, and thus the classification is not possible.

3.3 | Long-Term Foehn Reconstruction

Once the models from the previous section are estimated, they can be applied to the entire ERA5 period available. Although this is a prediction from a statistical perspective, it is termed 'reconstruction' in this article, as these predictions are applied backwards in time. The result is an hourly time series of foehn probabilities $\widehat{\mathrm{Pr}}_{1\mathrm{h}}$ from January 1, 1940 to December 31, 2022 (83 years).

This high-resolution reconstruction can serve as input for a variety of applications and analyses. To demonstrate the potential, we analyse the foehn occurrence from a climatological perspective: Did the occurrence of foehn increase/decrease along with the changing climate over the decades? Are there changes in the seasonal or diurnal patterns? These questions are investigated in more detail in the next section.

3.4 | Season-Trend Decomposition

The comprehensive reconstructed data set allows for the study of foehn occurrence in a climatological context. For this analysis, the hourly probability (Equation 3) is aggregated by (i) taking the highest probability $\widehat{\mathrm{Pr}}_{1\mathrm{h}}$ per day (0000 UTC–0000 UTC), before (ii) calculating monthly averages. The resulting time series contains "monthly means of the daily maxima", which are then modelled using a season-trend decomposition.

Due to the nature of the data, there is a large year-by-year but also within-year variability depending on the prevailing weather situation. To decompose the signal, a season-trend decomposition is employed separating the signal into long-term changes and a remainder component containing the residual variability.

In this study, the regression-based decomposition of Dokumentov and Hyndman (2022) is used, which also provides confidence intervals for the estimated season and trend components. The model for the monthly mean foehn probabilities y_t assumes an additive decomposition into a smoothly changing long-term trend T_t, a smoothly changing seasonal component $S_t^{(m)}$ and a remainder R_t:

$$y_t = T_t + \sum_{m=1}^{12} S_t^{(m)} + R_t \quad \text{with} \quad t \in 1, \ \ldots, J \tag{4}$$

where $m \in \{1, 12\}$ is the seasonal frequency (i.e., monthly) and J is the sample size (83 years × 12 months = 996). The model is estimated via the *R* package `stR` (Dokumentov and Hyndman 2023).

4 | Results

First, this section investigates which insights can be gained from the reconstruction about the foehn occurrence at the six different target stations. Different temporal scales are considered for this, namely: Inter-annual changes in the foehn probabilities in Section 4.1, long-term trends and seasonal patterns in Section 4.2, and changes in the diurnal patterns across decades in Section 4.3. All of these results are based on the reconstruction using the 'lasso' learner with the 'full' covariate set for the full-time period (without CV).

Second, the performance of the supervised learning model is assessed under different model specifications, namely: Using the 'full' set of all 497 variables versus the 155 'direct' variables only in Section 4.4 and comparing the performance of the three supervised learners (lasso, stability selection, extreme gradient boosting) in Section 4.5. All of these results are based on out-of-sample Brier scores (BSs) obtained in a six-fold CV.

4.1 | Average Annual Foehn Probabilities

The primary outcome of this study is the complete reconstruction of hourly foehn probabilities over 83 years (see Sections 3.1–3.3), yielding time series with $N \sim 7.27 \cdot 10^5$ individual probabilities $\widehat{\Pr}_{1h}$ (Equation 3). To carve out inter-annual variations in the foehn probabilities, we aggregate the reconstructed data by taking the daily maximum of $\widehat{\Pr}_{1h}$ before calculating annual means. This can be interpreted as the average probability of observing a foehn event on any given day within that year.

Figure 4 contains the result for all six stations, with Ellbögen exhibiting the highest mean annual probabilities (on average 0.334), while Altdorf and Innsbruck show the lowest (on average 0.130 and 0.133, respectively). Additionally, the annual mean of daily maxima from the classification is shown for years, with at least 80% of measurements available at the AWSs. The results show an overall good agreement between the two signals from the reconstruction and the classification, with some larger gaps due to data availability as well as some noticeable differences for specific stations in particular years.

The reconstruction reveals a pronounced inter-annual variability, with certain years exhibiting a much higher annual mean than the long-term average, whilst others distinctly fall below. This variability is not random, as one can see similar patterns among the four south-foehn stations. For instance, all four stations show unusually high mean probabilities for 1951 and 1972. Similarly, the two north-foehn stations exhibit a similar temporal behaviour over time.

Moreover, the figure suggests a possible increase in the annual mean foehn probability for south-foehn stations between 1940 and 1980. Hence, this question is investigated in more detail in the next section.

4.2 | Climatological Trends and Seasonal Patterns

In this section, the analysis from the previous section is taken a step further. Rather than focusing on the inter-annual variation the goal is to bring out the long-term climatological trends and changes in the seasonal patterns. Hence, the

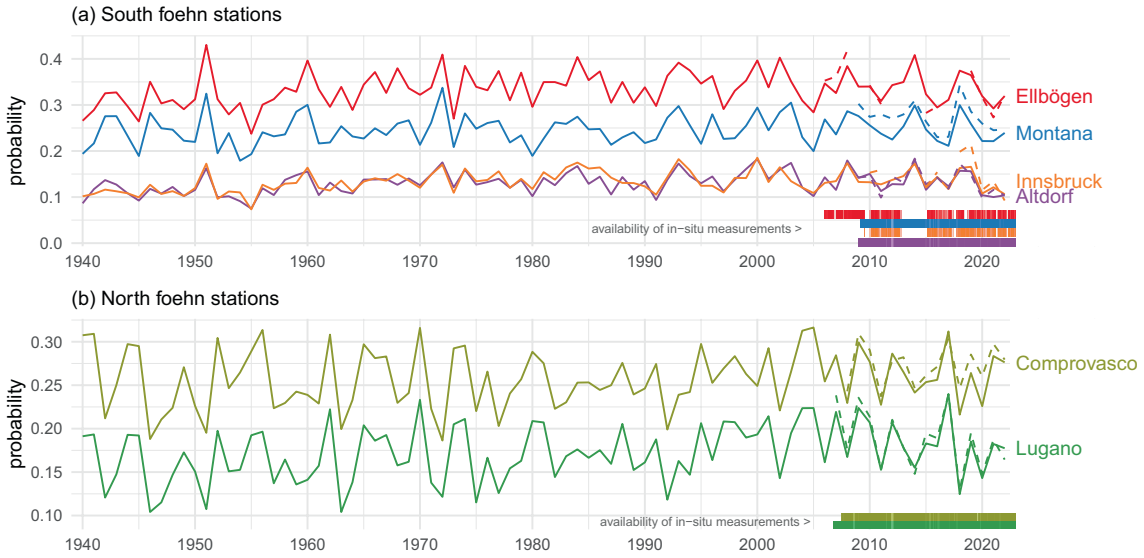

FIGURE 4 | Annual mean of highest daily foehn probability 1940–2020 (solid lines) for (a) the four south foehn and (b) the two noth foehn stations based on the 'lasso' model with 'full' covariate set. Additionally, annual means of daily maxima from the foehn classification using AWS data (dashed lines) and the availability of in situ measurements (straight bands at the bottom) are shown. Appendix S4 (Figure S1) shows a comparison to all other models and variants. [Colour figure can be viewed at wileyonlinelibrary.com]

reconstructed hourly time series are again aggregated but to monthly (rather than annual) means of the daily maxima of $\widehat{\mathrm{Pr}}_{1h}$. Based on the season-trend decomposition outlined in Section 3, Figure 5a illustrates the resulting smooth trends (T_t) along with the corresponding 95% confidence intervals and the long-term average. Figure 5b depicts the corresponding smoothly varying seasonal signals ($S_t^{(m)}$) averaged over decades for visual purposes (where the decade 1940 corresponds to the years 1940–1949 etc.).

The estimated trends (Figure 5a) show a slight increase across all six stations. At the two stations in Western Austria (Innsbruck, Ellbögen), an increase can be seen between 1940 and 1980, followed by a plateau. The other four stations exhibit a linear change over time. Although these changes are small in absolute terms, they are statistically significant for four of the six stations (all except Montana and Comprovasco). Here, significance indicates that the trend differs from a constant because the long-term average falls outside of the corresponding 95% confidence interval.

Figure 5b shows the analysis of the seasonal changes, which reveal the different characteristics between north-foehn stations and south-foehn stations. The two stations located south of the main Alpine ridge, Comprovasco and Lugano, show one distinct maximum in spring and a minimum during autumn. This pattern is stable over the entire study period, and no changes in the seasonal pattern ($S_t^{(m)}$) are found. The picture looks different when focusing on the four south-foehn stations which all show two maxima in spring and autumn with lower probability of foehn occurrence during summer and winter. Although not significant, the season-trend decomposition indicates an increase in the probability of foehn in spring (April, May) as well as in autumn (October, November) with a slight decrease in late summer (August, September).

4.3 | Diurnal Variability

For certain applications, information about diurnal patterns and their changes over time can be of great interest. With the hourly temporal resolution of the reconstruction, such insights are now possible across several decades.

Figure 6 shows Hovmøller diagrams for Ellbögen, depicting the decadal mean probability per time of day and month. Despite pronounced variability between the decades, the plot supports the previous findings showing an overall increase over time with the strongest increase in spring (April, May) and in autumn (October, November). In addition, this visualisation gives insights into the diurnal pattern. Generally, foehn occurrence in Ellbögen is more likely during the day (around 1000 UTC–2200 UTC) in spring and autumn, the period where the indication for a certain increase was found (Section 4.2). The minimum average foehn probability is in the early morning and tied to the length of the night. The time of the minimum varies somewhat interdecadally.

The Hovmøller diagrams for the remaining stations are provided in Appendix S5 (Figures S2–S6). Similar to Ellbögen, these diagrams support the findings from the previous sections while offering additional insights into the changes in diurnal patterns over the decades. For Comprovasco and Lugano, the diagrams show the same general patterns as the corresponding figures 20 and 37 from Cetti, Buzzi, and Sprenger (2015) for the time periods 1993–2003 and 2004–2014.

4.4 | Benefit of Full Covariate Set

As described in Section 3.2, two variants of the supervised learning methods are estimated: One using only the 155 'direct' variables and one using the 497 'full' variables, including large-scale atmospheric conditions (Section 2.2). For both variants, a six-fold CV is performed (Section 3.2).

To investigate the benefit of the 'full' covariate set, Figure 7 shows the BSs for the 'lasso' model for all six stations. For each station and each variant, BSs are shown for the test data set (out-of-sample) as well as for the training data set (in-sample). This shows that the models based on the 'full' variable set clearly outperform the models based on the 'direct' variables only. Although the overall performance of the less complex models based on the 'direct' variable set is still decent, including the additional large-scale spatio-temporal information substantially improves the overall model performance.

In addition to the predictive skill, Figure 7 also shows the stability of the model with the largest variance in the BSs visible on the test data sets in Ellbögen. On the training data sets, the scores barely vary due to the large sample size (10 years, hourly data). A comprehensive comparison of all models and variants can be found in Appendix S6 (Figure S7).

4.5 | Comparison of Supervised Learners

While the previous section focuses on the benefits of using more input data, this section compares the three different supervised learners described in Section 3.2. For simplicity, only the results for models based on the 'full' variable set are shown as they have been shown to outperform those only using the 'direct' variable set (Section 4.4).

Figure 8 illustrates that the BSs from the six-fold CV on the test set (out-of-sample) are comparable for all three supervised learners with only minor differences. The average BS is slightly lower for 'lasso' (0.0261), followed by 'stabsel' (0.0276) and 'xgboost' (0.0283).

On the training data (in-sample), the picture is similar for 'lasso' and 'stabsel', but 'xgboost' has much lower BSs. This indicates that 'xgboost' might be subject to some overfitting, despite careful tuning of the hyperparameters (Table S4 in Appendix S3.3).

5 | Discussion and Outlook

Using a novel combination of unsupervised and supervised learning, we are able to accurately reconstruct long-term foehn time series (starting from 1940) at hourly resolution. More specifically, foehn classification is first accomplished by (unsupervised)

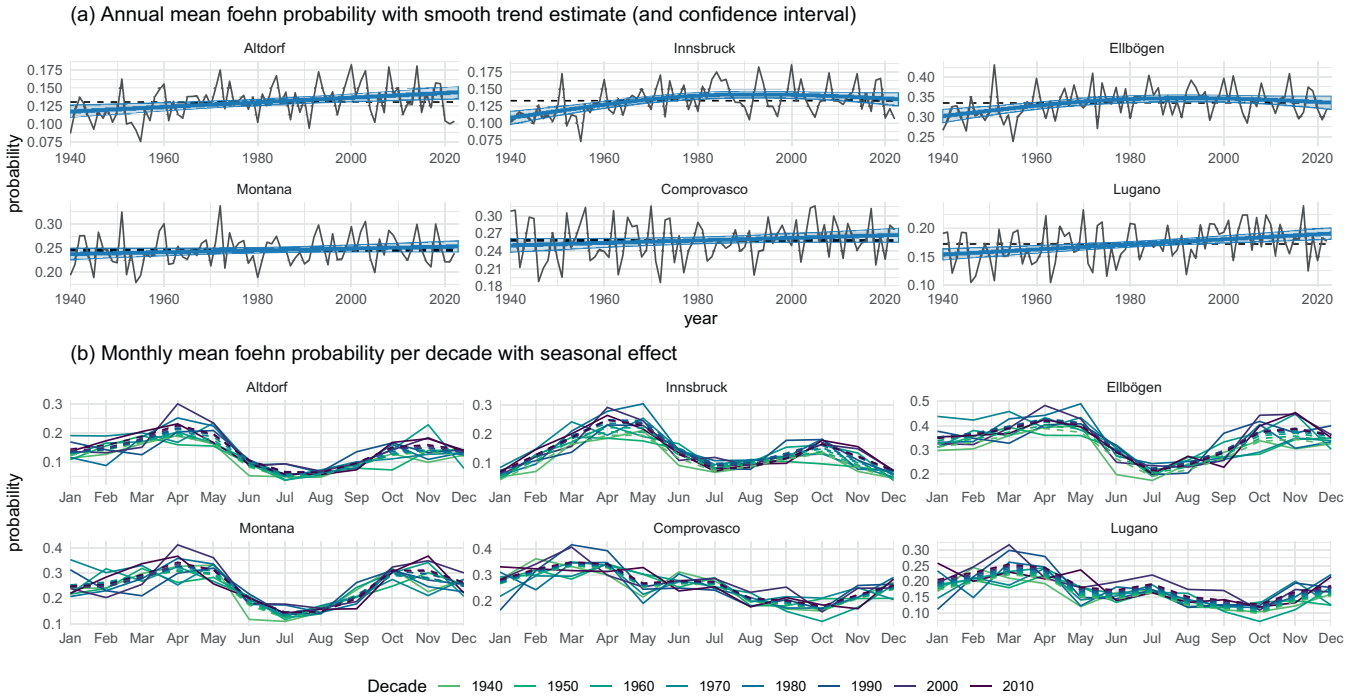

FIGURE 5 | Results of the season-trend decomposition. (a) Estimated long-term trends with estimated 95% confidence interval (blue) on top of annual mean probabilities (black), the horizontal dashed line shows the mean of the trend for better orientation. (b) Estimated seasonal pattern averaged over the whole period 1940–2022 (dashed) and separately by decade (solid, label indicates first year of each decade). [Colour figure can be viewed at wileyonlinelibrary.com]

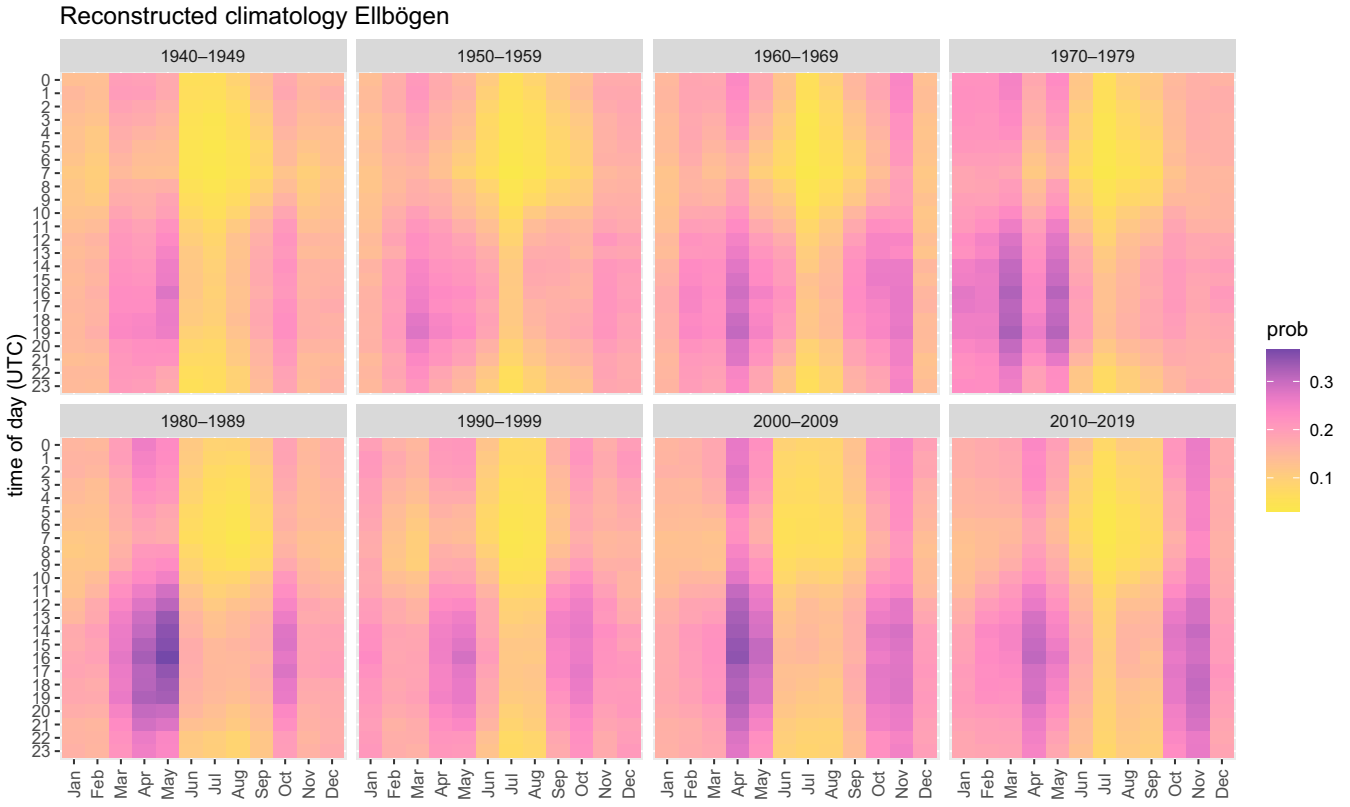

FIGURE 6 | Hovmøller diagrams showing the average foehn probability for Ellbögen over the course of the day (*y*-axis) for all months (*x*-axis) for the decades 1940–2019. [Colour figure can be viewed at wileyonlinelibrary.com]

Gaussian finite mixture models based on AWS measurements and then linked to ERA5 data using binary supervised learners such as lasso, stability selection or extreme gradient boosting. The resulting foehn reconstruction enables novel analyses, exemplified here by investigating long-term changes in trends, seasonal patterns and diurnal cycles of foehn occurrence.

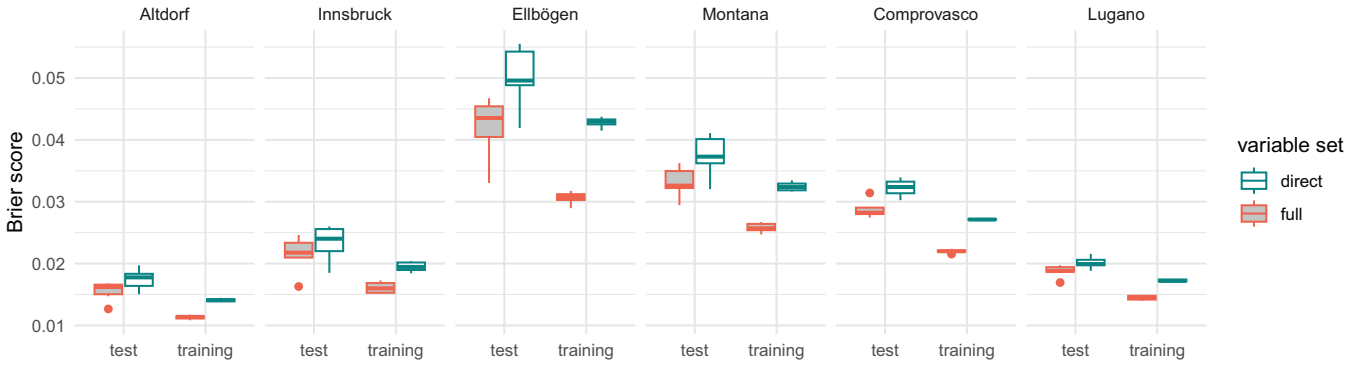

FIGURE 7 | Comparison of 'lasso' model performance using the 'direct' variable set (blueish) versus the larger 'full' variable set (orange with grey filling) for all stations. Shown is the average Brier score from the six-fold CV for the test (out-of-sample) and training (ins-sample) period including the years 2011–2022. Lower is better. [Colour figure can be viewed at wileyonlinelibrary.com]

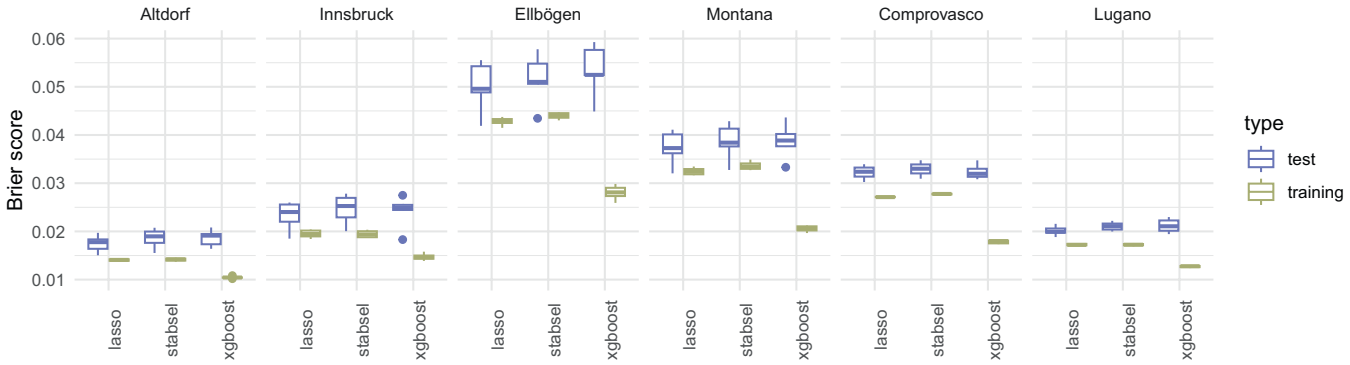

FIGURE 8 | Comparison of all supervised learning models using the 'full' variable set. Brier scores for training (in-sample; green) and test (out-of-sample; violet) data set based on six-fold cross-validation including the years 2011–2022. Lower is better. [Colour figure can be viewed at wileyonlinelibrary.com]

The season-trend decomposition based on the period 1940–2022 reveals that all six stations considered have either experienced a linear increase in foehn occurrence (probability) over the entire study period or an increase between 1940 and the early 1980s that levelled off afterwards. Altough these changes over time have proven to be statistically significant, they are small in absolute terms. The seasonality did not show any significant changes over time. However, the results for all south-foehn stations (Altdorf, Montana, Ellbögen and Innsbruck) indicate a slight increase in the occurrence of foehn in spring and autumn, with a slight decrease in late summer.

The high quality of the foehn reconstruction is partially due to using a large set of predictor covariates that not only contains information at the target location but also includes additional large-scale atmospheric information from the stations' surroundings. The benefits of this full set of covariates are similar for all three supervised learners considered: Logistic regression with lasso regularisation ('lasso'), logistic-regression-based stability selection ('stabsel') and extreme gradient boosting ('xgboost'). Lasso performs best in our application, closely followed by the other two learners. Some further improvements might be gained for xgboost if overfitting on the training data can be further reduced, e.g., by a different hyperparameter tuning strategy.

For a comparison to existing publications, we complement the BSs shown in the main paper by other popular scores for binary outcomes, such as the false negative rate (FNR, also known as miss rate), false positive rate (FPR, also known as false alarm rate) and percent correct (PC, also known as accuracy). Based on the best model (lasso with full covariate set), we obtain the following performances for Altdorf: 15.7%/0.4%/98.8% (FNR/FPR/PC). These align well with the existing literature and stand out for an exceptionally low FPR. Sprenger et al. (2017) report 11.8%/33.8%/96.5% and Mony, Jansing, and Sprenger (2021) report 21.4%/21.4%/97.7%. Similarly, for Lugano, we obtain 15.3%/0.7%/98.2%, while Mony, Jansing, and Sprenger (2021) report 22.1%/22.1%/97.1%. More details are included in the Appendix S6 (Table S5).

The same holds for the foehn classification when compared to existing literature and the Swiss foehn index (SFI) operationally used at MeteoSwiss in terms of 'average foehn hours per year'. The results from the Gaussian mixture model (Section 3.1) for Altdorf show an average of 482.4 h/year which aligns well with the SFI (458.4 h/year), as well as the results reported by Dürr (2008) (478 h/year), Jansing et al. (2022) (465.8 h/year) and MeteoSwiss (2024) (477 h/year). Montana exhibits 1007.4 h/year (SFI 904.7 h/year[1]), while Lugano and Comprovasco show 644.7 h/year (SFI 563.0 h/year, MeteoSwiss (2024) 551 h/year) and 1077.4 h/year (SFI 953.4 h/year), respectively.

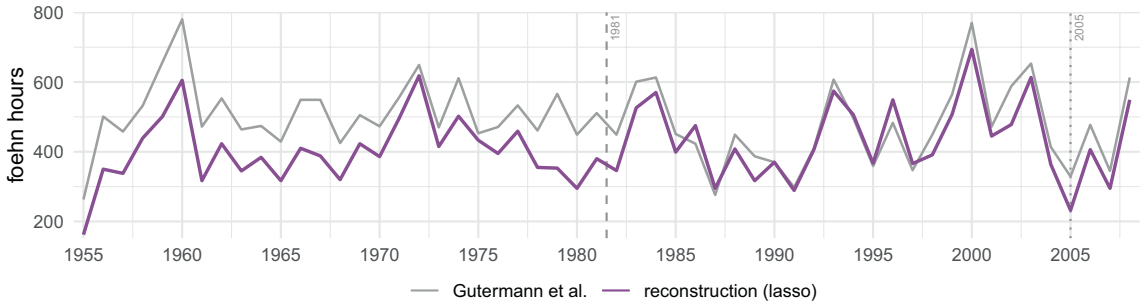

FIGURE 9 | Comparison of the annual number of foehn hours at Altdorf 1955–2008. Grey: Results by Gutermann and co-workers, using high-resolution AWS measurements starting in June 1981. Purple: Results obtained from our reconstruction with a training period starting in 2005. [Colour figure can be viewed at wileyonlinelibrary.com]

Lastly, the reconstruction for Altdorf is compared to the results from Gutermann et al. (2012) and Richner et al. (2014), who provide an hourly binary 'foehn'/'no foehn' time series for the period 1955–2008 at https://www.agfoehn.org. These data are aggregated to foehn hours per year and depicted in Figure 9 (grey) along with the annual number of foehn hours from our reconstruction ($\widehat{\Pr}_{1h} \geq 0.5$; purple). This shows that the results from both methods closely agree for the latter half of the time period. Only in the first half, the Guterman et al. series have systematically higher values than our reconstruction and the timing of the change essentially coincides with the availability of AWS measurements at Altdorf, starting in June 1981. Previously, the foehn indicators reported by Gutermann and co-workers are based on manual foehn classifications using traditional weather station recordings. Note that our reconstruction does not utilise any measurements from the station at Altdorf prior to the start of the training period in 2005.

It is worth mentioning that the quality of the reconstruction strongly relies on the quality of the automatic foehn classification. While the two-component Gaussian mixture model works well in this study, there are stations where this approach is insufficient. For example, we have found this to be the case for Aigle, Switzerland, where a three-component mixture model (or another classifier) appears to be necessary to separate light down-valley winds, strong humid (katabatic) outflows and actual foehn situations (details not shown).

However, if the separation of the AWS measurements into 'foehn' or 'no foehn' components works well (as for the six stations presented), the different binary supervised learners are able to link this reliably to the ERA5 data, yielding excellent results. In this article, we demonstrate the value of the long-term reconstruction by identifying long-term trends over the past decades as one possible application (Section 4.2). Additional work on this aspect might be needed in the future using different methods to get more detailed insights and assess the stability of our results.

Additionally, this high-resolution reconstruction offers promising opportunities for other applications. For instance, it can be used to fill gaps in historical records or to extend 'foehn observations' for studies, for example, studies which identify synoptic circulation patterns associated with foehn in specific regions (Kusaka et al. 2021; Stoev, Post, and Guerova 2022). Furthermore, these extended datasets can provide valuable insights into the effects of foehn on different areas, such as ecology, where the warming and drying effects of frequent foehn events could significantly impact flora and fauna or increase fire hazards.

It would also be interesting to see how the approach performs in other regions around the globe or when applied to the future (forecasts) rather than the past (reconstructions), similar to the work of Zweifel (2016), Sprenger et al. (2017) or Mony, Jansing, and Sprenger (2021). Although some adjustments will be needed, particularly for longer forecast horizons where the temporal resolution of NWP outputs typically decreases, this combination of supervised and unsupervised approaches has great potential for further research.

Author Contributions

Reto Stauffer: investigation, methodology, validation, visualization, writing – original draft, writing – review and editing, formal analysis, software, data curation, conceptualization. **Achim Zeileis:** writing – review and editing, data curation, conceptualization. **Georg J. Mayr:** writing – review and editing, data curation, conceptualization.

Acknowledgements

The study is partly based on the preliminary work of Morgenstern (2020). The computational results presented here have been achieved (in part) using the LEO HPC infrastructure of the University of Innsbruck.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

In situ observations were retrieved from the Swiss national weather service (MeteoSwiss; not publicly available) as well as Universität Innsbruck and GeoSphere Austria. The latter two are publicly available via https://acinn-data.uibk.ac.at/ and https://data.hub.geosphere.at/. ERA5 data is generated using Copernicus Climate Change Service publicly available via https://cds.climate.copernicus.eu/. Results for Ellbögen and Innsbruck are available at https://doi.org/10.48323/gdkr5-7tt45.

Computational Details

The results in this paper were obtained using *R* 4.2+. The majority of data preparation and handling is done using the *R* packages `stars` 0.6.4, `sf` 1.0.15 and `zoo` 1.8.12. `foehnix` 0.1.6 is used for foehn classification, and the supervised learning is based on `glmnet` 4.1.7 and

`xgboost` 1.7.5.1. The season-trend decomposition is based on the *R* package `stR` 0.6.

Endnotes

[1] Average based on the years 2009–2016 only.

References

Armi, L., and G. J. Mayr. 2007. "Continuously Stratified Flows Across an Alpine Crest With a Pass: Shallow and Deep Föhn." *Quarterly Journal of the Royal Meteorological Society* 133: 459–477.

Armi, L., and G. J. Mayr. 2011. "The Descending Stratified Flow and Internal Hydraulic Jump in the Lee of the Sierras." *Journal of Applied Meteorology and Climatology* 50: 1995–2011.

Armi, L., and G. J. Mayr. 2015. "Virtual and Real Topography for Flows Across Mountain Ranges." *Journal of Applied Meteorology and Climatology* 54: 723–731.

Cape, M. R., M. Vernet, P. Skvarca, S. Marinsek, T. Scambos, and E. Domack. 2015. "Foehn Winds Link Climate-Driven Warming to Ice Shelf Evolution in Antarctica." *Journal of Geophysical Research: Atmospheres* 120: 11037–11057.

Cetti, C., M. Buzzi, and M. Sprenger. 2015. "Climatology of Alpine North Foehn." *Scientific Report MeteoSwiss* 100: 76. https://www.meteoswiss.admin.ch/dam/jcr:816f0ea5-d1a8-4b3c-8cf9-753d370ad532/SR_Cetti.pdf.

Chen, T., and C. Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, 785–794. New York, NY, USA: Association for Computing Machinery.

Courvoisier, H., and T. Gutermann. 1971. "Zur Praktischen Anwendung Des Föhntests Von Widmer." http://www.agfoehn.org/doc/Courvoisier_1971.pdf.

Datta, R. T., M. Tedesco, X. Fettweis, et al. 2019. "The Effect of Foehn-Induced Surface Melt on Firn Evolution Over the Northeast Antarctic Peninsula." *Geophysical Research Letters* 46: 3822–3831.

Dokumentov, A., and R. J. Hyndman. 2022. "STR: Seasonal-Trend Decomposition Using Regression." *INFORMS Journal on Data Science* 1: 50–62.

Dokumentov, A., and R. J. Hyndman. 2023. "stR: STR Decomposition." https://CRAN.R-project.org/package=stR. R package version 0.6.

Dürr, B. 2008. "Automatisiertes Verfahren Zur Bestimmung Von Föhn in Alpentälern, Arbeitsbericht, MeteoSchweiz." https://www.meteoschweiz.admin.ch/dam/jcr:1fbe0d4e-fc7c-4121-b89e-1429a931272f/ab223.pdf. Accessed 2024-03-21.

Elvidge, A. D., P. Kuipers Munneke, J. C. King, I. A. Renfrew, and E. Gilbert. 2020. "Atmospheric Drivers of Melt on Larsen C Ice Shelf: Surface Energy Budget Regimes and the Impact of Foehn." *Journal of Geophysical Research: Atmospheres* 125: e2020JD032463.

Francis, D., R. Fonseca, K. S. Mattingly, S. Lhermitte, and C. Walker. 2023. "Foehn Winds at Pine Island Glacier and Their Role in Ice Changes." *Cryosphere* 17: 3041–3062.

Friedman, J. H., T. Hastie, and R. Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33: 1–22.

Grajek, Z., and E. Bednorz. 2024. "Climatology and Circulation Conditions of Potential Foehn Occurrence in the Polish Tatra Mountains." *Acta Geophysica*.

Grisogono, B., and D. BelušIć. 2009. "A Review of Recent Advances in Understanding the Mesoand Microscale Properties of the Severe Bora Wind." *Tellus A: Dynamic Meteorology and Oceanography* 61: 1–16.

Grün, B., and F. Leisch. 2008. "FlexMix Version 2: Finite Mixtures With Concomitant Variables and Varying and Constant Parameters." *Journal of Statistical Software* 28: 1–35.

Gutermann, T., B. Dürr, H. Richner, and S. Bader. 2012. *Föhnklimatologie Altdorf: Die Lange Reihe (1864–2008) Und Ihre Weiterführung, Vergleich Mit Anderen Stationen*. Vol. 241, 1–53. Fachbericht MeteoSchweiz. https://www.meteoswiss.admin.ch/services-and-publications/publications/reports-and-bulletins/2012/foehnklimatologie-altdorf--die-lange-reihe--1864-2008--und-ihre-.html.

Hersbach, H., B. Bell, P. Berrisford, et al. 2023a. "ERA5 Hourly Data on Pressure Levels From 1940 to Present."

Hersbach, H., B. Bell, P. Berrisford, et al. 2023b. "ERA5 Hourly Data on Single Levels From 1940 to Present."

Jansing, L., L. Papritz, B. Dürr, D. Gerstgrasser, and M. Sprenger. 2022. "Classification of Alpine South Foehn Based on 5 Years of Kilometre-Scale Analysis Data." *Weather and Climate Dynamics* 3: 1113–1138.

Koyanagi, T., and H. Kusaka. 2020. "A Climatological Study of the Strongest Local Winds of Japan "Inami-Kaze"." *International Journal of Climatology* 40: 1007–1021.

Kusaka, H., A. Nishi, A. Kakinuma, Q.-V. Doan, T. Onodera, and S. Endo. 2021. "Japan's South Foehn on the Toyama Plain: Dynamical or Thermodynamical Mechanisms?" *International Journal of Climatology* 41: 5350–5367.

Laffin, M. K., C. S. Zender, S. Singh, J. M. Van Wessem, P. Smeets, and C. H. Reijmer. 2021. "Climatology and Evolution of the Antarctic Aeninsula Föhn Wind-Induced Melt Regime From 1979–2018." *Journal of Geophysical Research: Atmospheres* 126: 1–19.

Mayr, G. J., and L. Armi. 2008. "Föhn as a Response to Changing Upstream and Downstream Air Masses." *Quarterly Journal of the Royal Meteorological Society* 134: 1357–1369.

Mayr, G. J., and L. Armi. 2010. "The Influence of Downstream Diurnal Heating on the Descent of Flow Across the Sierras." *Journal of Applied Meteorology and Climatology* 49: 1906–1912.

McClung, B., and C. F. Mass. 2020. "The Strong, Dry Winds of Central and Northern California: Climatology and Synoptic Evolution." *Weather and Forecasting* 35: 2163–2178.

McGowan, H. A., and A. P. Sturman. 1996. "Regional and Local Scale Characteristics of Foehn Wind Events Over the South Island of New Zealand." *Meteorology and Atmospheric Physics* 58: 151–164.

Meinshausen, N., and P. Bühlmann. 2010. "Stability Selection." *Journal of the Royal Statistical Society B* 72: 417–473.

MeteoSwiss. 2024. "Wetter Und Klima A Bis Z: Föhnhäufigkeit." https://www.meteoschweiz.admin.ch/wetter/wetter-und-klima-von-a-bis-z/foehnhaeufigkeit.html. Accessed 2024-03-21.

Mony, C., L. Jansing, and M. Sprenger. 2021. "Evaluating Foehn Occurrence in a Changing Climate Based on Reanalysis and Climate Model Data Using Machine Learning." *Weather and Forecasting* 36: 2039–2055.

Morgenstern, D. S. 2020. "Multidecadal Foehn Time Series Reconstruction Using Machine Learning and ERA5 Reanalysis Data. Master's thesis, Universität Innsbruck." https://bibsearch.uibk.ac.at/AC15677542.

Muñoz, R. C., and L. A. Armi. 2024. "Hydraulic Analysis and Cold-Air Pool Interaction of the Raco Gap Wind." *Journal of Applied Meteorology and Climatology* 63: 505–526.

Norte, F. A. 2015. "Understanding and Forecasting Zonda Wind (Andean Foehn) in Argentina: A Review." *Atmospheric and Climate Sciences* 5: 163–193.

Plavcan, D., and G. J. Mayr. 2015. "Towards an Alpine Foehn Climatology." https://www.uibk.ac.at/congress/icam2015/abstracts_

oral_presentations.htm#O14.4, 33rd International Conference on Alpine Meteorology, Innsbruck,Accessed 2024-03-21.

Plavcan, D., G. J. Mayr, and A. Zeileis. 2014. "Automatic and Probabilistic Foehn Diagnosis With a Statistical Mixture Model." *Journal of Applied Meteorology and Climatology* 53: 652–659.

Reinhard, M., M. Rebetez, and R. Schlaepfer. 2005. "Recent Climate Change: Rethinking Drought in the Context of Forest Fire Research in Ticino, South of Switzerland." *Theoretical and Applied Climatology* 82: 17–25.

Richner, H., B. Dürr, T. Gutermann, and S. Bader. 2014. "The Use of Automatic Station Data for Continuing the Long Time Series (1864 to 2008) of Foehn in Altdorf." *Meteorologische Zeitschrift* 23: 159–166.

Richner, H., and P. Hächler. 2013. "Understanding and Forecasting Alpine Foehn, Chap." 4, 219–260, Dordrecht: Springer-Verlag.

Rolinski, T., S. B. Capps, and W. Zhuang. 2019. "Santa Ana Winds: A Descriptive Climatology." *Weather and Forecasting* 34: 257–275.

Schoennagel, T., T. T. Veblen, and W. H. Romme. 2004. "The Interaction of Fire, Fuels, and Climate Across Rocky Mountain Forests." *Bioscience* 54: 661–676.

Sergius, L. A., G. R. Ellis, and R. M. Ogden. 1962. "The Santa Ana Winds of Southern California." *Weatherwise* 15: 102–121.

Śliwińska, M., and D. Ciaranek. 2015. "Very Strong Foehn Winds in the Tatra Mountains (Polish Carpathian Mountains): Causes, Course and Consequences." *Aerul şi Apa. Componente Ale Mediului; Air and Water Components of the Environment* 2015: 109–116.

Speirs, J. C., H. A. McGowan, D. F. Steinhoff, and D. H. Bromwich. 2013. "Regional Climate Variability Driven by Foehn Winds in the McMurdo Dry Valleys, Antarctica." *International Journal of Climatology* 33: 945–958.

Sprenger, M., S. Schemm, R. Oechslin, and J. Jenkner. 2017. "Nowcasting Foehn Wind Events Using the AdaBoost Machine Learning Algorithm." *Weather and Forecasting* 32: 1079–1099.

Stauffer, R. 2023. "Foehnix: A Toolbox for Automated Foehn Classification Based on Mixture Models." https://retostauffer.github.io/Rfoehnix/. R package version 0.1.6.

Stoev, K., P. Post, and G. Guerova. 2022. "Synoptic Circulation Patterns Associated With Foehn Days in Sofia in the Period 1979–2014." *IDŐJÁRÁS/Quarterly Journal of the Hungarian Meteorological Servise* 126: 545–566.

Tay, J. K., B. Narasimhan, and T. Hastie. 2023. "Elastic Net Regularization Paths for all Generalized Linear Models." *Journal of Statistical Software* 106: 1–31.

Turton, J. V., A. Kirchgaessner, A. N. Ross, and J. C. King. 2018. "The Spatial Distribution and Temporal Variability of Föhn Winds Over the Larsen C Ice Shelf, Antarctica." *Quarterly Journal of the Royal Meteorological Society* 144: 1169–1178.

Wessen, J. M. V., and M. K. Laffin. 2022. "Regional Atmospheric Climate Model 2 (Racmo2), Version 2.3p2." Accessed 2023-08-23.

Widmer, R. 1966. "Statistische Untersuchungen Über Den Föhn Im Reusstal Und Versuch Einer Objektiven Föhnprognose Für Die Station Altdorf, Vierteljahresschrift Der Naturforschenden Gesellschaft Zürich." https://www.ngzh.ch/archiv/1966_111/111_3-4/111_20.pdf, Accessed 2024-03-21, 111, 331–375.

Zumbrunnen, T., H. Bugmann, M. Conedera, and M. Bürgi. 2009. "Linking Forest Fire Regimes and Climate–A Historical Analysis in a Dry Inner Alpine Valley." *Ecosystems* 12: 73–86.

Zweifel, L. 2016. "Probabilistic Foehn Forecasting for the Gotthard Region Based on Model Output Statistics. Master's Thesis, Universität Innsbruck." https://permalink.obvsg.at/UIB/AC11359704.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.

Article XIII

Fix F., Mayr G. J., Zeileis A., Stucke I.K., and Stauffer R. (2024). *Atmospheric Deserts: Detection and Consequences*, Weather and Climate Dynamics, *forthcoming*, doi:10.5194/egusphere-2024-2143.

JCR ranking: **Category 1** in *Meteorology & Atmospheric Sciences*.

Contribution (CRT): *Data curation / software / writing, review and editing*.

https://doi.org/10.5194/egusphere-2024-2143
Preprint. Discussion started: 17 July 2024

Atmospheric Deserts: Detection and Consequences

Fiona Fix[1], Georg Mayr[1], Achim Zeileis[2], Isabell Stucke[1], and Reto Stauffer[3]

[1]Department of Atmospheric and Cryospheric Sciences, Universität Innsbruck
[2]Department of Statistics, Universität Innsbruck
[3]Department of Statistics & Digital Science Center, Universität Innsbruck

Correspondence: Fiona Fix (fiona.fix@uibk.ac.at)

Abstract. We introduce the concept of atmospheric deserts (ADs), air masses that are advected away from hot and dry convective boundary layers in semi-arid or desert source regions. They can be expected to eliminate cloudiness, cause heat to build up in the target region, suppress thunderstorm formation in their centre and boost thunderstorm formation at their edges. A direct detection method, tracing the AD from source to target using Lagrangian trajectories is developed.

We illustrate this new concept of ADs with a case study in Europe from mid-June 2022. With the Lagrangian analysis tool (LAGRANTO) approximately 200 million trajectories are calculated, tracking the path of the air mass and the development of its properties as it progresses from North Africa towards and across Europe over the course of five days. k-means-clustering identifies four typical pathways that the trajectories follow. For one of the pathways, the air nearly conserves its well-mixed properties. Diabatic processes of radiative cooling, latent heating due to condensation, and cooling due to re-evaporation of precipitation, however, modify the air along the other pathways.

The case study demonstrates how ADs influence the weather in the target region. Thunderstorms are mainly absent in the centre of the AD, but erupt along a line parallel to its boundary. At this edge of the AD and the surface front, lifting occurs, causing the formation of thunderstorms. The AD does not reside directly above the local boundary layer for long enough to be the main cause for the heat wave affecting large parts of Europe, but may contribute to it. Subsidence heating of another air stream, was identified as one possible reason for the increased near-surface temperatures.

1 Introduction

Severe thunderstorms and heat waves bear serious risks for human health, economy, and society. Heat waves are the reason for many deaths (e.g., Schär and Jendritzky, 2004; Schär, 2016) especially in highly populated regions like Europe, and thunderstorms can cause severe (economic, ecologic, ...) damage. Understanding weather situations that influence these extreme weather events is therefore of great importance. We postulate that atmospheric deserts (ADs) can greatly impact heat wave and thunderstorm formation. This paper introduces ADs as air masses that originate in the hot and dry convective boundary layers (CBLs) of semi-arid, desert, subtropical and/or elevated source regions. They can strongly influence the vertical temperature and moisture profiles in the regions they are advected into and create large temperature and moisture gradients at their lateral boundaries.

https://doi.org/10.5194/egusphere-2024-2143
Preprint. Discussion started: 17 July 2024

Previous research has studied a small subset of possible manifestations of ADs by looking for well-mixed, warm, and dry layers on top of the local boundary layer (BL) in vertical profiles in the target region (e.g. Carlson and Ludlam, 1968; Carlson et al., 1983; Lanicci and Warner, 1991a; Banacos and Ekster, 2010; Cordeira et al., 2017; Ribeiro and Bosart, 2018), sometimes complemented by numerical weather prediction models (Arritt et al., 1992) or satellite imagery (Gitro et al., 2019). These layers are termed "elevated mixed layers" (EMLs, e.g. Carlson et al., 1983; Banacos and Ekster, 2010; Ribeiro and Bosart, 2018, and others). They occur in the special case where the properties of the AD remain (almost) constant during the advection. More commonly, however, diabatic processes will modify the ADs along their way. These modifications together with differential advection in the vertical will make the air mass unrecognizable in the vertical profiles in the target region. We therefore expect EMLs to be (much) rarer than their generalization, the ADs.

The effects on the weather in the target region should, however, be similar (evidence was found by Johns and Dorr, 1996), although possibly of smaller magnitude. These effects stem from strong temperature and moisture gradients at the vertical and lateral boundaries of ADs (e.g. Carlson et al., 1983; Farrell and Carlson, 1989; Dahl and Fischer, 2016). For the special case of EMLs it was found that the hot and dry air masses ride up over the cooler, moister, shallower CBL in the target region, and can form a capping inversion (or "lid", e.g., Carlson and Ludlam, 1968; Carlson, 1980; Carlson et al., 1983; Lanicci and Warner, 1991a, b; Cordeira et al., 2017). The lid can lead to heat buildup underneath, especially under the typically associated cloud free conditions, which leads to an increase in the convective available potential temperature (CAPE) and the near-surface temperatures (e.g., Carlson and Ludlam, 1968; Carlson, 1980; Carlson et al., 1983; Keyser and Carlson, 1984; Farrell and Carlson, 1989; Cordeira et al., 2017). It was found that (severe) thunderstorms often erupt along the edge of the EML (Carlson and Ludlam, 1968; Carlson et al., 1980, 1983; Keyser and Carlson, 1984; Lanicci and Warner, 1991c; Arritt et al., 1992; Banacos and Ekster, 2010; Lewis and Gray, 2010; Sibley, 2012; Dahl and Fischer, 2016; Cordeira et al., 2017, and others). Towards the edge of the EML, the lid base height increases and its strength decreases, hence the constraint on convection at the edge is not as strong as in the central area, where the lid suppresses thunderstorm eruption (Carlson et al., 1983).

Since more than one third of the Earth's land surface is arid (e.g. Vaughn, 2005; Tchakerian, 2015; European Commission et al., 2018), ADs might play an important, yet understudied role in mid-latitude weather. This study introduces the new concept of ADs and a direct way to identify them and trace their properties along their way. It illustrates the impact of an AD originating in North Africa on the weather in Europe over a five-day period in June 2022.

2 Definition of atmospheric deserts and detection method

ADs are air masses that are advected away from the hot and dry CBL of semi-arid, desert, subtropical and/or elevated source regions. These air masses progressively lose their distinct characteristics during the advection over hundreds to thousands of kilometers due to diabatic processes and differential advection in the vertical. A special case of an AD is an EML, in which case the air mass remains (almost) unmodified and well-mixed.

As ADs are generally modified during the advection, indirect detection methods based on the properties in the target region are ambivalent and often insufficient. Therefore, we introduce a novel detection method, that traces the air mass directly from

https://doi.org/10.5194/egusphere-2024-2143
Preprint. Discussion started: 17 July 2024

its source to the target region, using Lagrangian trajectories. The AD is then defined as all the grid boxes (any grid chosen for the application at hand) that contain at least one trajectory at a given time and the development of the AD along its path can be analyzed. The detection method requires a trajectory calculation tool and a spatio-temporally complete data set of atmospheric data over a large area covering the source and the target regions. A high vertical resolution of the meteorological data is crucial for the detection and analysis of the ADs.

2.1 Trajectory calculation

The AD air is traced directly from the source to the target region. Hence, forward trajectories need to be initiated in the source region continuously. Per definition, the origin of an AD is the CBL of an arid source region. In these regions, the depth of the CBL can reach up to several kilometres during daytime, especially in the summer months (Garcia-Carreras et al., 2015). However, as also the diurnal cycle can be strong, during nighttime CBLs can be very shallow. Nevertheless, air parcels with "boundary layer-like" thermodynamic properties reside in the residual layer (former daytime CBL). There is no physical reason why those air parcels should not be advected together with those from the CBL, and therefore contribute to the AD. Since the residual layer cannot be easily determined and is often not available in data sets, a simple approach can be used to approximate it. The actual boundary layer height (BLH) is smoothed using *splines*. This results in the smoothed BLH (BLH_s) which is equal to the actual BLH during the day (i.e. from 13:00 to 17:00 UTC), and smoothed in between (an explanatory figure can be found in Appendix A, Fig. A1). In order to detect and analyse the AD, trajectories are hence initiated from below the smoothed BLH, BLH_s, in the source region, and meteorological variables are traced along them. The spatio-temporal resolution of the initialization, the length of the trajectories, and the size of the grid boxes depend on the application and the available data.

2.2 Trajectory clustering

In order to simplify the analysis of very many trajectories, we group them into several clusters, representing *typical trajectory pathways*. Identifying typical pathways requires defining the characteristics for comparison. Here, we consider spatial (longitude, latitude, altitude), thermodynamic, and microphysical aspects. We employ the 11 variables listed in Tab. 1 and the data driven k-means-clustering method (MacQueen, 1967) to cluster the trajectories. A similar approach was also used by Nie and Sun (2022). This method identifies k clusters (groups) in the data, where the data points within one cluster are as similar as possible, while the clusters are as dissimilar as possible. The measure for the similarity is the squared Euclidean distance between each data point and the cluster means.

To capture the spatial aspects, the differences between the final and starting positions of the trajectories are calculated for longitude, latitude, and height above mean sea level ($hamsl$): $\mathrm{diff}_\mathrm{lon}$, $\mathrm{diff}_\mathrm{lat}$, and diff_{hamsl}. Additionally, the maximum difference in $hamsl$ in all 6 h windows along the trajectory ($\mathrm{diff}_{hamsl,\mathrm{max}}$) is useful to distinguish between trajectories that rise slowly and those that rise abruptly. To account for subsidence after an initial ascent, we introduce the difference between the maximum and the final height, $\mathrm{diff}_{hamsl,\mathrm{subs}}$. Since different diabatic processes such as radiation, mixing, or condensation/evaporation have different impacts on the changes of the thermodynamic and microphysical variables, the differences of the potential temperature θ (diff_θ), the specific humidity q (diff_q) and the total (logarithmic) water content cwc (diff_{cwc}) between

https://doi.org/10.5194/egusphere-2024-2143
Preprint. Discussion started: 17 July 2024

Variable	Formula
diff_{lon}	lon at the final location – lon at the starting point
diff_{lat}	lat at the final location – lat at the starting point
diff_{hamsl}	$hamsl$ at the final location – $hamsl$ at the starting point
$\text{diff}_{hamsl,\text{max}}$	maximum(6-hourly differences in $hamsl$)
$\text{diff}_{hamsl,\text{subs}}$	maximum($hamsl$) – $hamsl$ at final location
diff_{θ}	θ at the final location – θ at the starting point
diff_{q}	q at the final location – q at the starting point
diff_{cwc}	(ln($ciwc$) + ln($cswc$) + ln($clwc$) + ln($crwc$)) at the final location
	– (ln($ciwc$) + ln($cswc$) + ln($clwc$) + ln($crwc$)) at the starting point
$\text{corr}_{\theta,q}$	correlation(θ, q)
$\text{corr}_{\text{diff}_{\theta},ccwc}$	correlation(6-hourly differences of θ, ln($ciwc$)) + correlation(6-hourly differences of θ, ln($clwc$))
$\text{corr}_{\text{diff}_{\theta},cpwc}$	correlation(6-hourly differences of θ, ln($cswc$)) + correlation(6-hourly differences of θ, ln($crwc$))

Table 1. Variables characterizing the trajectories, used in k-means-clustering. For further explanation regarding their meaning, refer to the text.

the final and starting locations are taken into account. Here, cwc is the sum of the four types of cloud water content: liquid ($clwc$), ice ($ciwc$), rain ($crwc$), and snow ($cswc$). Cloud water content variable distributions are typically heavily skewed towards low values. This skewness can be reduced by performing a logarithmic transformation. The diabatic processes do not only affect the differences in the thermodynamic and microphysical variables, but also on their evolution. Some processes lead to correlated changes in temperature and moisture, for example, while for other processes the changes are unrelated. Hence, the correlation between the potential temperature and the specific humidity ($\text{corr}_{\theta,q}$) and the correlations between the precipitating and non-precipitating cloud water contents and the 6-hourly differences in the potential temperature are taken into account. Here, we refer to the precipitating cloud water content ($cpwc$) as the sum of $crwc$ and $cswc$, and the non-precipitation cloud water content ($ccwc$) as the sum of $clwc$ and $ciwc$. The respective correlations are denoted as $\text{corr}_{\text{diff}_{\theta},cpwc}$ and $\text{corr}_{\text{diff}_{\theta},ccwc}$. All these variables are summarized in Tab. 1 and need to be normalized by subtracting their mean and dividing by their standard deviation before applying the clustering method, so that those with large values do not dominate the clustering.

The cluster averages can then be used for analyzing the development of the properties along the different *typical pathways*. The mean is calculated in the trajectory relative time frame (hours since initialization), rather than in the absolute time frame, since trajectories typically move faster than the synoptic situation, so that this yields a more comparable result. Note that this average is not a trajectory itself and might not be physically consistent across variables. An exemplary figure and a more detailed description of the effects of averaging can be found in Appendix A (Fig. A2).

https://doi.org/10.5194/egusphere-2024-2143
Preprint. Discussion started: 17 July 2024

3 Practical Application

3.1 Data and trajectory model

The model chosen to calculate the trajectories in this work is the Lagrangian analysis tool (LAGRANTO) version 2.0, which has been developed since the late nineties (Sprenger and Wernli, 2015), and is therefore a mature and widely used tool in the atmospheric sciences, in various contexts (e.g. Stohl et al., 2001; van der Does et al., 2018; Keune et al., 2022; Oertel et al., 2023). Forward or backward trajectories can be calculated iteratively based on the 3D wind field of the input dataset with customized starting region and resolution. In this study, Europe is the target region, hence North Africa is used as the source region. A polygon marking the source region can be seen as the grey outline in Fig. 1. Earlier studies took the Iberian peninsula as the source region for European EMLs (e.g. Carlson and Ludlam, 1968; Karyampudi and Carlson, 1988; Lewis and Gray, 2010; Dahl and Fischer, 2016), however we find that including Iberia only increases the number of trajectories by a few percent and the influences on the results are marginal. Trajectories are started at a very high resolution of 5 km in the horizontal and 10 hPa in the vertical between 1100 and 400 hPa, from below BLH_{s}.

As a spatio-temporally complete data set of atmospheric data, we use the latest global reanalysis from the European Centre for Medium-Range Weather Forecasts (ECMWF), ERA5 (Hersbach et al., 2020), which is based on the Integrated Forecasting System (IFS) Cy41r2. The horizontal resolution of ERA5 is 0.25°, and data are available hourly on 137 vertical model levels up to 1 Pa (Hersbach et al., 2020). This results in a high vertical resolution of about 20 m at the surface and 300 m at 500 hPa. The domain chosen for this study covers Northern Africa and Europe, specifically 30° W to 60° E and 15° to 73° N. ERA5 single-level, pressure-level and model-level data on the lowest 74 model levels are obtained (surface to about 120 hPa). Additional to the reanalysis variables, we obtain the mean temperature tendency due to short and long wave radiation ($mttswr, mttlwr$), which are only available as forecast variables. We use the forecast from 06:00 UTC for the times from 09:00 UTC to 20:00 UTC and the forecast from 18:00 UTC for the times between 21:00 UTC and 08:00 UTC to avoid the respective spin-up periods. According to the IFS documentation (European Centre for Medium-Range Weather Forecasts, 2016), the different types of water content can be converted into one another by condensation, melting, autoconversion, etc. The rain, snow, and ice particles ($crwc, cswc, ciwc$), are allowed to sediment. Constant fall speeds of 4, 1, and 0.13 ms^{-1} are assumed, respectively. Precipitation is allowed to be advected by the 3D-wind, and to re-evaporate when falling through an environment with lower relative humidity than a critical value. For more detail the reader is referred to the model documentation (European Centre for Medium-Range Weather Forecasts, 2016; Hersbach et al., 2020).

The trajectories are aggregated to grid boxes of 0.25° times 0.25° times 500 m, matching ERA5 grid cells in the horizontal. An AD-cell is a grid box that contains at least one trajectory. Lightning is used as a proxy for thunderstorm location, hence a lightning measurement data set is suitable to analyse the connection between ADs and convection. This study uses data from the lightning location network “Blitzortung" (Wanke et al., 2014). The network processes data from sensors operating at the very low frequency range set up by a large number of volunteers around the world. In this frequency range, weaker strokes are not detected and detection efficiency varies slightly between day and night. However, the network still allows to reliably detect

https://doi.org/10.5194/egusphere-2024-2143
Preprint. Discussion started: 17 July 2024

locations of more widespread lightning activity. Consequently, those records where only a single flash was observed within a radius of one ERA5 grid cell were omitted.

3.2 Exemplary case study

A case study in June 2022 is used as an example to explain the concept of ADs and first findings in this study. Central Europe experienced intense heat during the period between 18 June 2002 and 19 June 2022 (e.g. Imbery et al., 2022, also see Fig. 3). Intense heat was recorded in southern Europe already in the days prior. Low pressure systems were located so that advection of air from Northern Africa to Europe occurred.

The five-day period from 15 to 19 June 2022 is chosen for this case study. Trajectories are initiated hourly between 15 June 2022 00:00 UTC and 19 June 2022 11:00 UTC. This results in approximately 200 million trajectories starting from the smoothed North African BL during this case study. Of these, only about 37 million pass north of 37° N at least once and have not left the domain yet by 19 June 2022 12 UTC and are therefore interesting for further analysis.

As the synoptic situation changes during the period of the case study, trajectories initiated at different times will follow different pathways. However, we do not expect big differences during one day, hence we chose to apply the clustering on all trajectories that start on the same day. We standardize the variables and apply the clustering method explained in Sect. 2.2 on each initialization day separately. 18.8/13.2/4.8/0.2 million trajectories are clustered for the initialization days 15/16/17/18 June 2022, respectively (of those trajectories starting on 19 June 2022, none cross 37° N before 12:00 UTC). In this case, four clusters are chosen, as the total sum of squares does not decrease drastically for more clusters. Note that due to the different lengths of the trajectories, during the last 23 h since initialization, fewer trajectories are averaged, when calculating the cluster average. Of all the grid boxes north of 37° N that were identified as AD cells at 19 June 2022 12 UTC, the respective clusters C1 to C4 (of all days combined) cover 43.4 %, 63.9 %, 26.0 %, and 29.3 %, respectively.

Figure 1 shows the evolution of the geopotential, AD maximal extent, AD extent at 800 hPa, and 800 hPa fronts during the period 16–19 June 2022. A low pressure system, initially located south west of the Iberian peninsula, moves north east, across the Gulf of Biscay during the case study. It is with the warm air sector of this low pressure system, that air from North Africa is advected northwards. Another low pressure system travels from south of Greenland to southwestern Scandinavia, its cold front is shown in blue in Fig. 1. About 42 % of all columns north of 37° N (within 30° W to 60° E and 37° to 73° N) are covered by the AD for at least for one hour during the study period. By the end of the period shown, large parts of Europe are covered by the AD (see Fig. 1, at June 19 2022 12 UTC, 29 % of all columns within the domain 30° W to 60° E and 37° to 73° N are covered by the AD), which extends as far north as the British Isles and southern Sweden and as far east as Russia, north of the Black Sea. AD-cells can be found at any altitude between the ground and up to 13 km. The black outline of the maximum extent of the AD in Fig. 1 shows that while the entire AD still resides south of the 800 hPa cold front at 18 June 2022 00:00 UTC, its edge passes north of it during the next day. However, the edge of the AD at the 800 hPa level remains south of the cold front for longer, and the front only catches up with it in parts by the end of the case study. The lateral boundary of the AD is also marked by a strong horizontal temperature gradient that sharpens over time but remains separate from synoptic frontal boundaries.

https://doi.org/10.5194/egusphere-2024-2143
Preprint. Discussion started: 17 July 2024

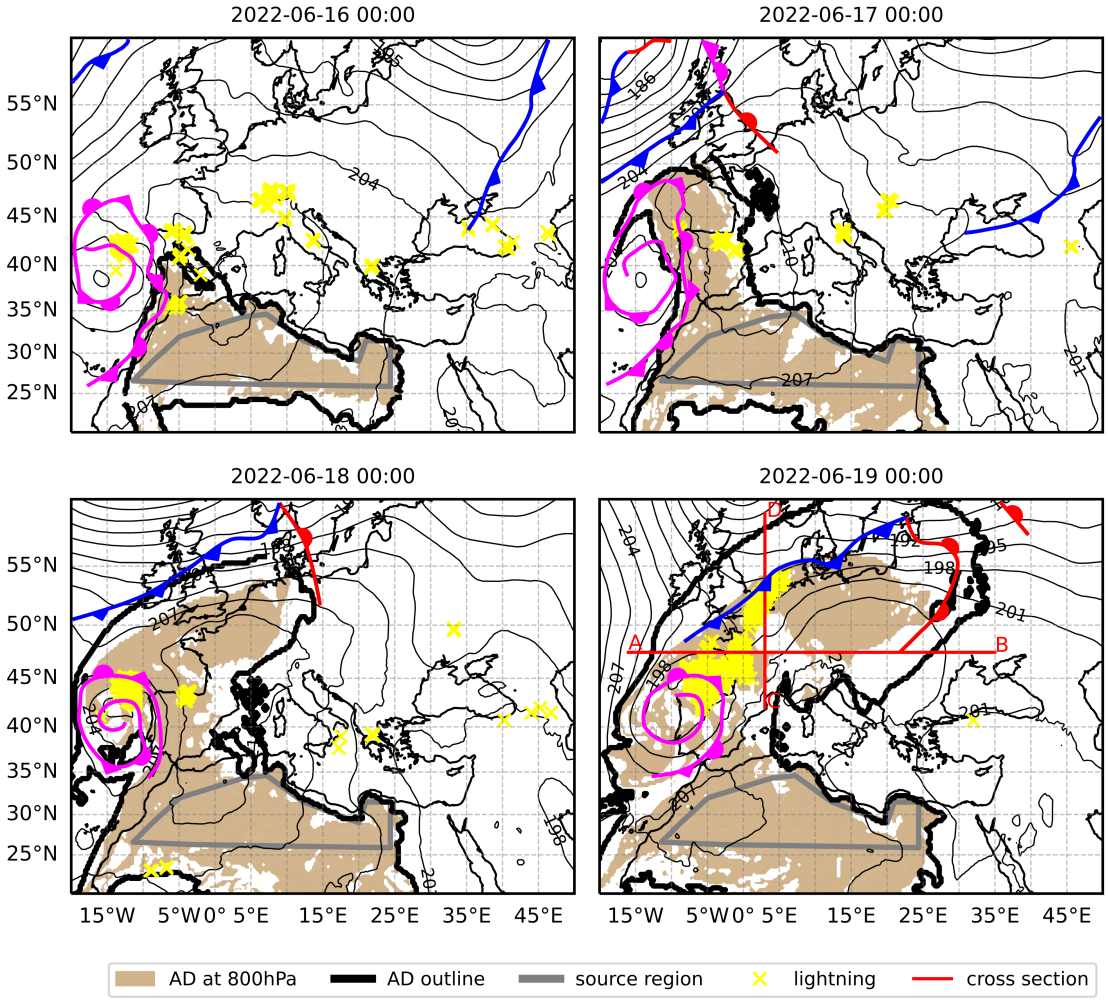

Figure 1. Display of the situation during the second half of the case study, 16–19 June 2022, 00:00 UTC, respectively. Thin black contours show the 800 hPa geopotential height in decametres, with a spacing of 3 dam. The coloured lines denote the 800 hPa fronts, colours and symbols have their usual meaning. The maximum extent of the AD is outlined in thick black. The extent of the AD in the layer from 800 to 750 hPa is marked in beige. Yellow crosses mark locations where lightning occurred during the hour before and after. Red lines (A–B) and (C–D) mark the locations of the cross-sections depicted in Figs. 4 and 5.

Yellow crosses in Fig. 1 mark locations where lightning occurred in the 2 h window centred at the given time. Thunderstorms persist in the vicinity of the occluding front west of the Iberian peninsula and in the Gulf of Biscay. Lightning also occurs along the edge of the AD. Especially during the night of 19 June 2022, there is a distinctive line of lightning close to the edge of the 800 hPa AD-layer. Notably, most of the area covered by the AD, especially in its centre, does not experience any thunderstorm activity. This is analyzed more closely in Sect. 3.4.3.

https://doi.org/10.5194/egusphere-2024-2143
Preprint. Discussion started: 17 July 2024

3.3 Modification of the thermodynamic properties

The cluster averages for the 18.78 million trajectories started on 15 June 2022 are shown in Fig. 2. Averages are calculated over 6.1, 6.7, 3.6, and 2.4 million trajectories for clusters C1, C2, C3, and C4, respectively. The figure shows the development of different thermodynamic variables with time since initialization (panels a–f), as well as the path across the map (Fig. 2g).

During the first 24 to 36 h, all four clusters behave similarly and rise from approximately 2 km to approximately 4 km (Fig. 2a), while their potential temperature increases (Fig. 2b), and the absolute temperature decreases (Fig. 2c). During this time the trajectories still reside in North Africa (Fig. 2g). The cumulative mean temperature tendency due to radiation (Fig. 2b, dashed) does not explain this warming, and the specific humidity does not indicate that latent heating is responsible, hence it can be assumed that mixing with a warmer air mass, such as the local CBL, is responsible for this increase in potential temperature.

After these initial 24 to 36 h the clusters begin to differ: Clusters C1, C2, and C3 (cyan, blue, and red) follow a very similar geographical path, crossing the Iberian peninsula, turning east over the Gulf of Biscay and travelling further east across Northern France and Germany towards Eastern Europe (Fig. 2g). Cluster C1 follows what could be called the typical or expected behaviour of an EML. It rises slightly (Fig. 2a) while riding up on local air masses, but it almost conserves its potential temperature (Fig. 2b) and specific water vapour content (Fig. 2d), which would preserve the well-mixed properties of the North African CBL.

Diabatic processes, however, modify the properties of the trajectories in the other clusters. The trajectories in Cluster C2 (blue) rise much higher on average (up to approximately 8 km, Fig. 2a). An especially sudden ascent is visible around hour 80. During this ascent, the trajectories cool adiabatically (Fig. 2c), which induces condensation (decrease in q, Fig. 2d, and increase in cloud water content variables, Fig. 2e,f). Latent heat causes the potential temperature in this cluster to rise (Fig. 2b) and condensate precipitates out (the total water content q_t decreases, not shown here, as it is almost indistinguishable from q in Fig. 2d). This sudden ascent coincides with the location where the cluster rises on top of the cooler air mass in the north (Figs. 1 and 2g). With this behaviour, this cluster is the most similar to a warm conveyor belt (e.g. Browning, 1971).

In contrast, cluster C3 (red) remains at a constant height above mean sea level after the ascent during the initial 24 h and then experiences a descent around hour 80. Meanwhile, its potential temperature decreases (Fig. 2b) and its specific water vapour content increases (Fig. 2d). This is partly due to radiative cooling (dashed in Fig. 2b), and partly due to evaporative cooling as precipitation falling through from above re-evaporates. This explanation is supported by the fact that together with the decrease in potential temperature (Fig. 2b), the specific water vapour content increases (Fig. 2d). Re-evaporation is possible in the data used here, since ice, snow and rain are allowed to sediment and can re-evaporate when they fall through a sub-saturated air mass in ERA5 (European Centre for Medium-Range Weather Forecasts, 2016). The strong correlation between the precipitation cloud water contents of C2 and C3 (dashed in Fig. 2e, f) together with the increase of specific water content (Fig. 2d) gives trust in this explanation. Also, mixing with the cooler, moister local air can be a reason for the cooling and moistening. As the cluster average is comprised of many different trajectories, it is likely that all three processes play a role.

The trajectories in the fourth cluster (C4, orange) take longer to leave North Africa and turn east already over the Iberian peninsula on average, so that they almost reach the Mediterranean Sea. A closer analysis of the trajectories in this cluster shows

https://doi.org/10.5194/egusphere-2024-2143
Preprint. Discussion started: 17 July 2024

however, that this cluster is rather heterogeneous and is comprised of trajectories that turn east early and ones that are led around the low pressure system to the west counter-clockwise. As those trajectories likely also experience different processes altering their properties, this cluster is more difficult to interpret than the others, which are more homogeneous. Additionally, these trajectories mainly reside over the Gulf of Biscay and the Mediterranean, therefore, they are less important to interpret in the context of the AD's consequences for central Europe.

The cluster paths and the development of the thermodynamic variables along the path are similar across the first three initialization days in the case study (see Supplementary Material), only the average path of the trajectories started on June 18 differ considerably as expected for the short travelling time. It can be observed that on the later days, the trajectories of clusters C1, C2, and C3 reach further north on average, which makes sense considering that the guiding low pressure system has moved in between. The trajectories started on June 18 experience a similar increase in height and potential temperature as discussed here for the initial 12 h, but then they turn east earlier, travelling along the southern Spanish coast and hence never meet the colder air mass in the north.

3.4 Consequences for the local weather in the target region

Warm air advection aloft can suppress cloud formation by confining the local convection to a shallow layer below, hence ADs should be associated with cloud free conditions in the target region. Indeed, large parts of the area covered by the AD are cloud free or only covered by high clouds during the entire study period (not shown here). Medium and low clouds preceding the 800 hPa cold front indicate the region where the AD rises up on the colder air mass at its northwestern edge. It becomes apparent from Fig. 1 that the 800 hPa cold front approaches the AD from the northwest, but only catches up (in parts) with the AD at this level by the end of the case study. This is in accordance with the findings by Dahl and Fischer (2016), who find a similar behaviour in their 3-year composite analysis of EMLs with convergence lines.

Figs. 4 and 5 show vertical cross-sections along 47.5° N (A–B) and 3° E (C–D), for 19 June 2022 00:00 UTC, as marked with red lines in Figs. 1 and 3. The AD (grey contour/grid) covers large parts of the lower and middle troposphere, resides higher aloft at its edges and even comes as far down as the surface in its centre, hence penetrating the local BL during the day (not shown here, but discussed in more detail in Sect. 3.4.2). It is not well-mixed in terms of potential temperature (Figs. 4a and 5a) or total water content (not shown here). Hence, the AD does not classify as an EML and could not be identified from vertical profiles in the target region, which highlights the necessity of the direct detection method presented in this study for the analysis of ADs. The cold front is clearly visible in the potential temperature in both cross-sections (marked in blue). The horizontal temperature gradient at the western edge of the AD (Fig. 4a) is comparable in strength to the one at the cold front. Even more pronounced it is seen in the equivalent potential temperature due to increased humidity (Fig 4b,c). Similarly, the horizontal and vertical gradients in potential (and equivalent potential) temperature at the northern edge of the AD (Fig. 5a,b) are of similar magnitude as those at the cold front. At the southern edge the gradients are also visible, but weaker. Hence, the lateral edges of the AD are strongly baroclinic zones.

Above the cold air mass at the surface, there is another air mass present, which is clearly drier (panel (c) in Figs. 4 and 5). This air mass is separated by strong gradients in the (equivalent) potential temperature from the cold air below and the AD

https://doi.org/10.5194/egusphere-2024-2143
Preprint. Discussion started: 17 July 2024

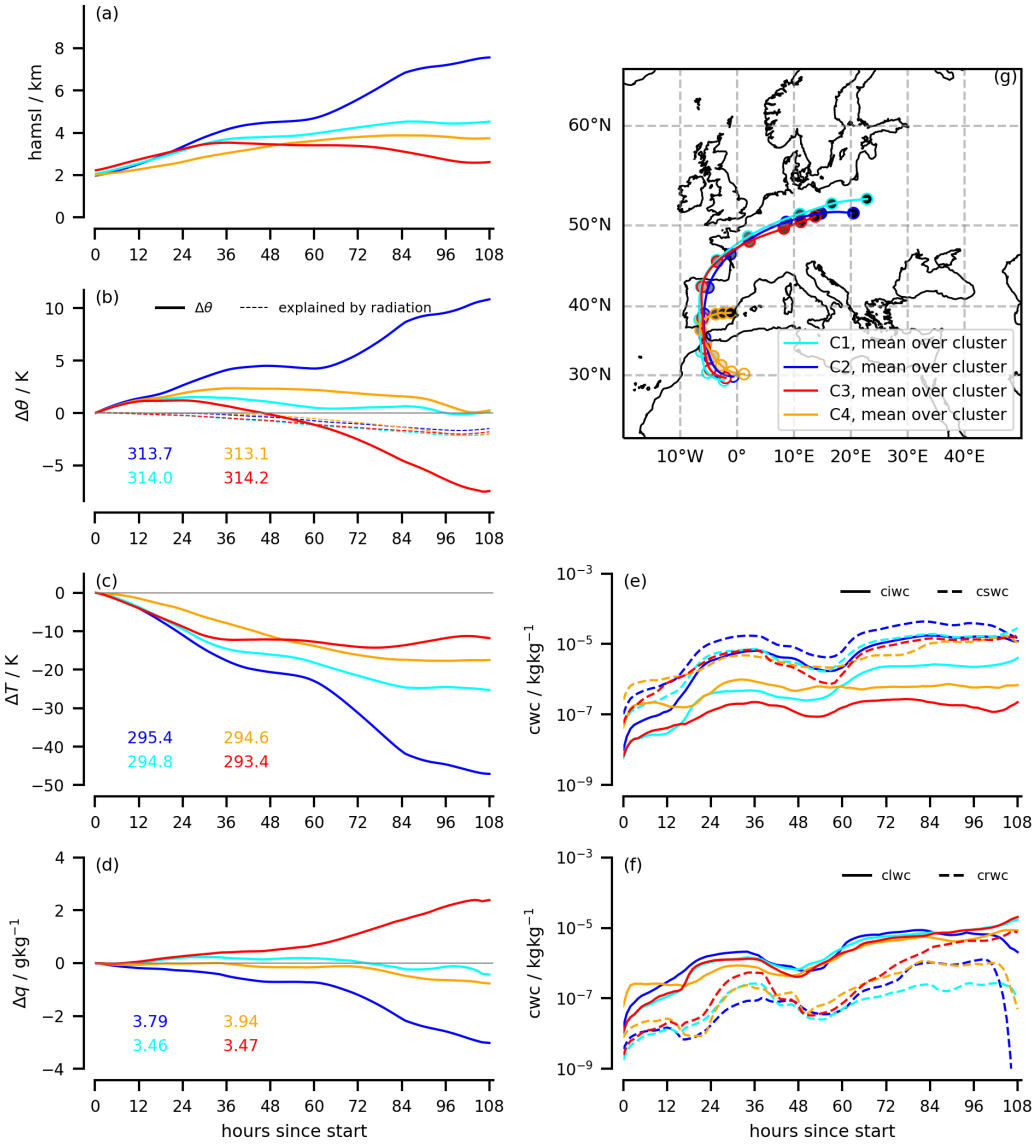

Figure 2. Time series of thermodynamic variables (panels (a–f)) and map of the cluster paths (g) for the cluster mean for C1 (cyan), C2 (blue), C3 (red), and C4 (orange). The mean is calculated in the trajectory relative time frame, the x-axis of the time series shows hours since the trajectory initialization. Panel (a): height above mean sea level ($hamsl$). Panel (b): Difference in potential temperature (θ) since initialization and cumulative mean temperature tendency due to short- and longwave radiation (dashed). Panel (c): Difference in temperature (T) since initialization. Panel (d): Difference in specific water content (q). Panel (e): cloud ice water content ($ciwc$, solid) and cloud snow water content ($cswc$, dashed). Panel (f): cloud liquid water content ($clwc$, solid) and cloud rain water content ($crwc$, dashed). Panel (g) shows a map of the mean trajectory path. Dots mark every 12th one-hour time step (which corresponds to the x-ticks in the other panels), the colour gradient of the dots represents progression in time, with white being the time of initialization. The coloured numbers is panels (b–d) denote the initial value of the respective variables and clusters in K (b,c) and gkg^{-1} (d). Panels (e–f) have a logarithmic scale on the y-axis.

https://doi.org/10.5194/egusphere-2024-2143
Preprint. Discussion started: 17 July 2024

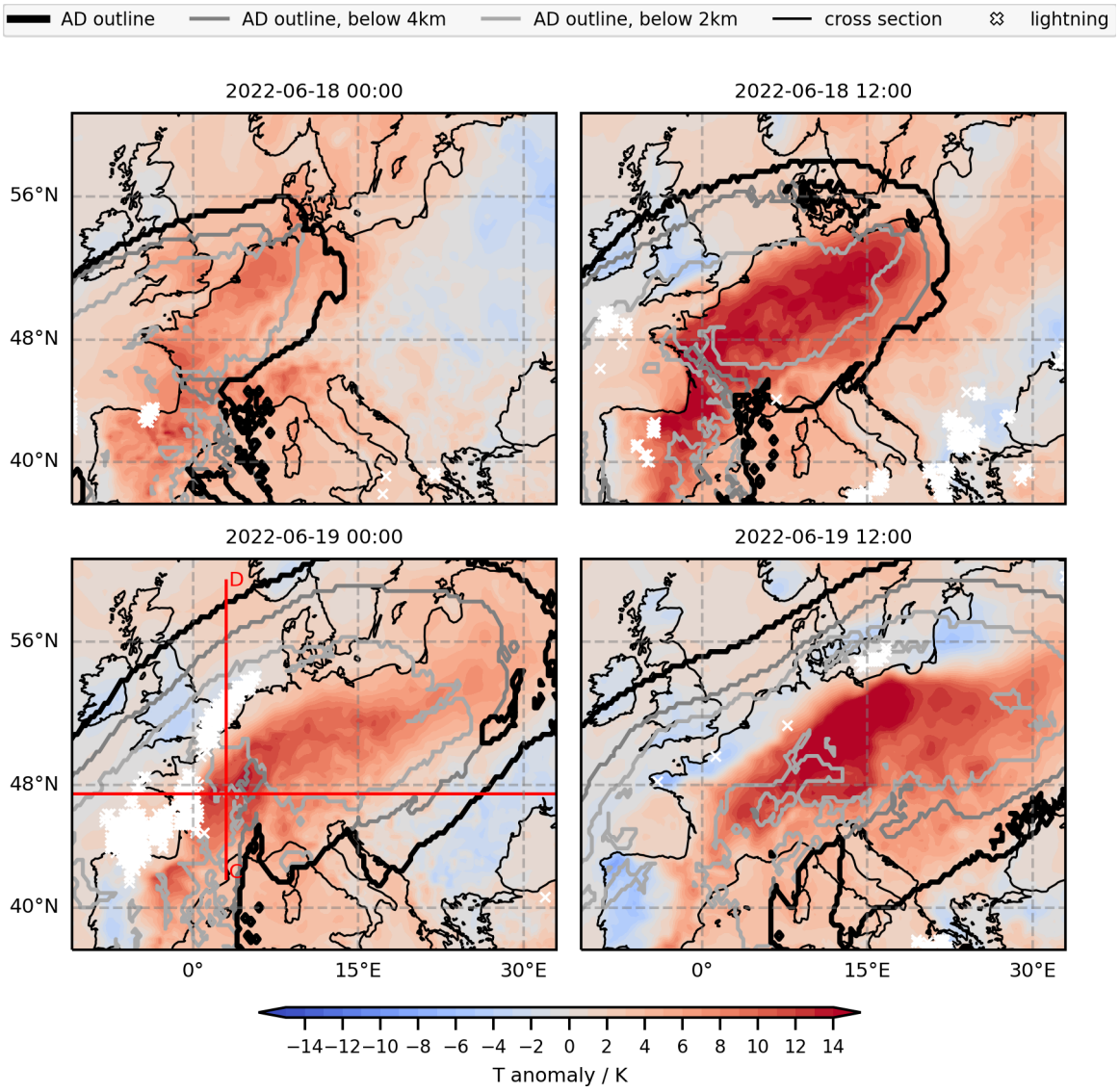

Figure 3. Map showing the spatial extent of the AD and the 2 m temperature anomaly with respect to the 30 year period 1992–2021 at 06:00 UTC and 12:00 UTC on 18 and 19 June 2022. The entire AD is outlined in black, outlines of the AD cells up to 4 and 2 km, respectively, are marked in grey. White crosses mark locations where lightning occurred during the hour before and after, 2 m temperature anomalies are coloured with 2 K spacing.

https://doi.org/10.5194/egusphere-2024-2143
Preprint. Discussion started: 17 July 2024

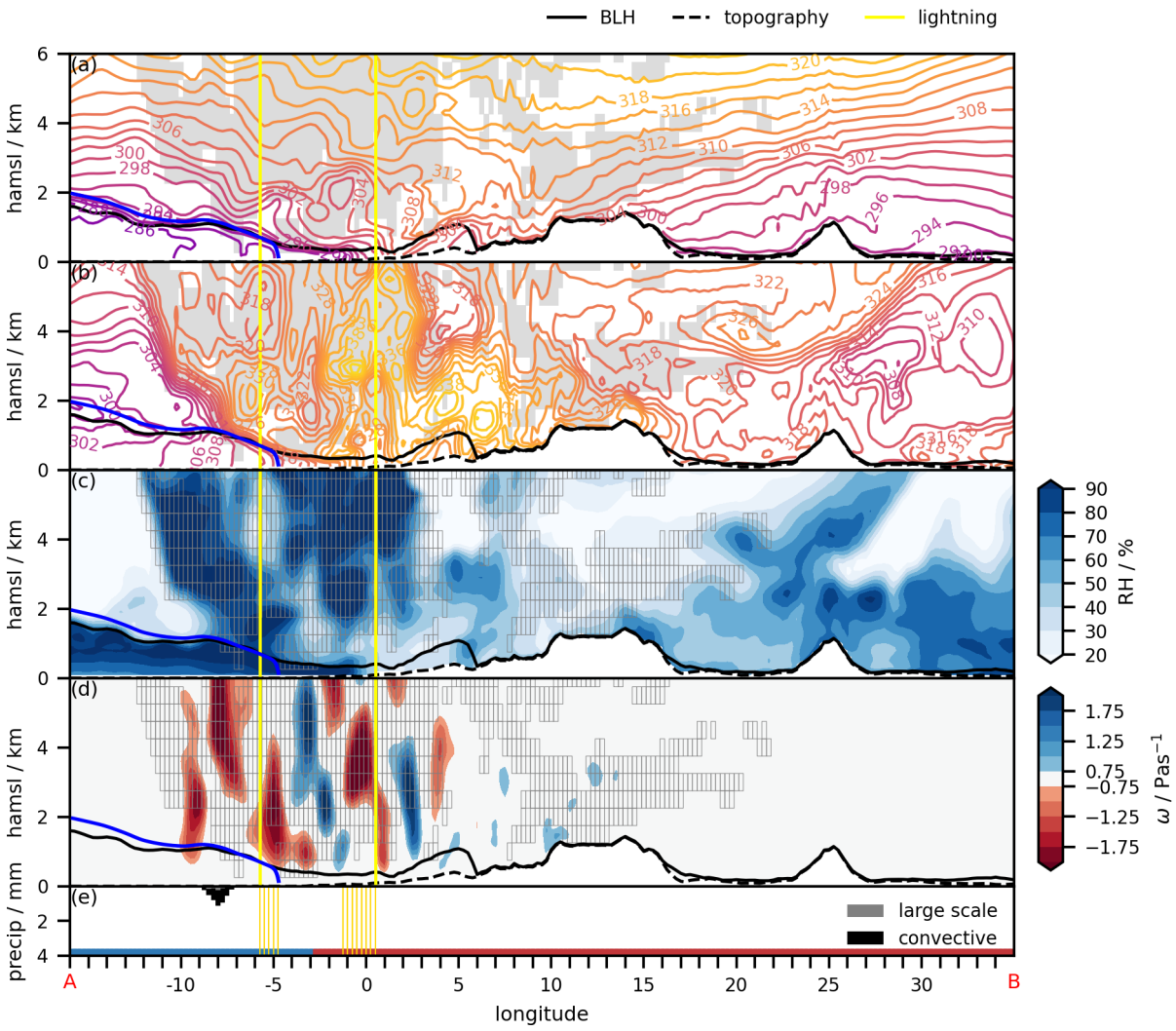

Figure 4. Vertical cross-section along 47.5° N, at 19 June 2022 00:00 UTC (16° W to 35° E, 47.5° N, as denoted by the red line (A–B) in lower left panel in Fig. 1). Shown are (a) the potential temperature in K, (b) the equivalent potential temperature, (c) the relative humidity in %, (d) the vertical wind component in Pas^{-1}, and (e) the accumulated large scale (gray) and convective (black) precipitation within the previous hour in mm. In panels (a–d), the solid black line denotes the BLH, the dashed black line the model topography, and the vertical yellow lines denote the range in which lightning occurred within a 2 h time window centred at 00:00 UTC and a 1° latitude band centred at 47.5° N. All lightning locations within this range are shown in yellow in panel (e). The region occupied by AD air is marked in grey (shading in (a, b), grid in (c, d)). The cold front is denoted in blue. Land and ocean surfaces are marked along the x-axis in brown and blue, respectively.

https://doi.org/10.5194/egusphere-2024-2143
Preprint. Discussion started: 17 July 2024

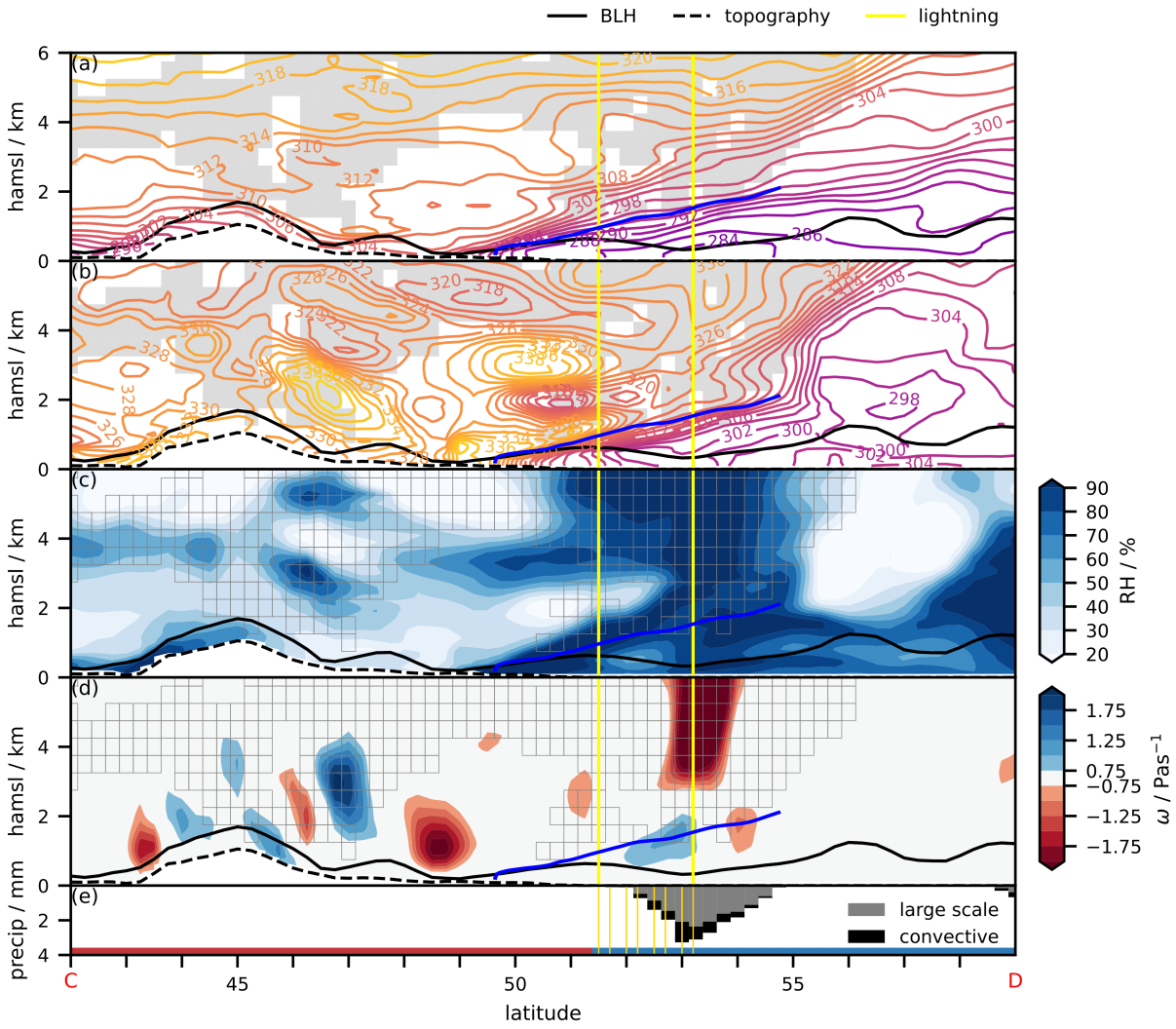

Figure 5. Vertical cross-section along 3° E, at 19 June 2022 00:00 UTC (3° E, 42° N to 59° N, as denoted by red line (C–D) in lower left panel in Fig. 1). As Fig. 4.

https://doi.org/10.5194/egusphere-2024-2143
Preprint. Discussion started: 17 July 2024

above. The edge of the AD is visible even more clearly in the gradient of the equivalent potential temperature (Figs. 4b and 5b).

It was suggested that the presence of EMLs can lead to heat waves (e.g. Cordeira et al., 2017). The increased near-surface temperatures seen in Fig. 3 (2 m temperature anomalies compared to 1992–2021) are further discussed in Sect. 3.4.1. The formation of thunderstorms is also influenced by the presence or absence of a lid and is further discussed in Sect. 3.4.3.

3.4.1 High near-surface temperatures

In large parts of Europe, the surface temperatures were exceptionally high during the AD event presented here (e.g. Imbery et al., 2022, and Fig. 3). It was proposed that the warm air of EMLs aloft form capping inversions due to their high potential temperatures, which prevent the local BL from growing and reduce vertical mixing (e.g. Carlson and Ludlam, 1968; Carlson et al., 1983; Farrell and Carlson, 1989; Cordeira et al., 2017). Especially under clear-sky conditions, this allows the surface temperatures, equivalent potential temperatures, and CAPE to rise to exceptionally high values and lead to heat waves (Cordeira et al., 2017).

The AD may form a lid when the lowest AD-cell in a column lies just above the local BL. At 19 June 2022 12 UTC, 20 % of all the AD-columns north of 37° N have a "lid", identified by the centre of the lowest AD-cell in the column being within ±500 m of the ERA5 BLH. However, the lid was not present for long enough to cause the high near-surface temperatures. Only in 0.34 % of all the AD-columns north of 37° N a lid was present for longer than 24 h. Most of these columns are in the vicinity of the occluding front in the Gulf of Biscay and over the ocean.

Other possible explanation for the high surface temperatures are advection or subsidence heating. Hence, back-trajectories from the BL in two regions in eastern Germany and southwestern France which experienced exceptionally high temperatures (Imbery et al., 2022) were calculated (see Fig. 6). Some of the back-trajectories from the local BL in those regions, indeed originate in western North Africa and are therefore part of the AD, which penetrated the local BL (see discussion in Sect. 3.4.2). A larger part of the back-trajectories originates over the Atlantic and travels across central Europe, where it subsides. Temperatures increase due to adiabatic heating during the descent. Hence, the analysis of the AD event in June 2022 supports the hypothesis that ADs co-occur with heat waves. However, in this case the AD did not cause the heat wave.

3.4.2 Penetration of AD-air into the target region's boundary layer

In the centre of the AD air mass, AD air penetrates the local BL. The impact of the trajectories that enter the local BL on the clustering analysis is small, they constitute only 1.1 % of the 37 million trajectories in total. For the AD air to be entrained into the local BL, they must have a similar potential temperature. As already mentioned, the near-surface temperature was already unusually high. Additionally, the trajectories that reside within the local BL by 19 June 2022 12:00 UTC, have cooled considerably (about 6 K, for those trajectories initiated on 15 June 2022, while the trajectories that end up above the local BL actually warm (2 K on average for those initiated on 15 June 2022). For one, the trajectories penetrating the local BL experience less radiative cooling on average (not shown here). Second, the development of the specific water content indicates that the warming is due to latent heating due to condensation, while the cooling is due to evaporation, likely of precipitation falling

https://doi.org/10.5194/egusphere-2024-2143
Preprint. Discussion started: 17 July 2024

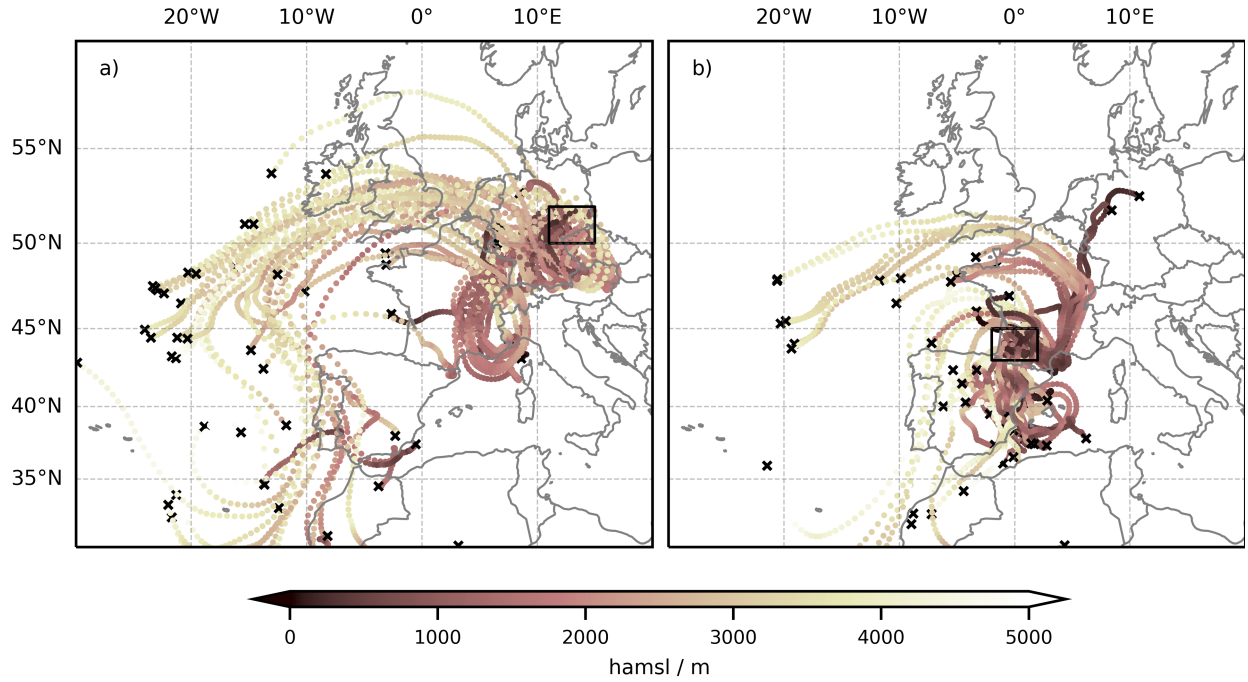

Figure 6. Back-trajectories from areas with elevated surface temperatures. A random subset of 20 back-trajectories that are within the local BL at the time and location of initiation is shown. Colours indicate the height above mean sea level ($hamsl$) in metres. Black crosses mark the position at 15 June 2022 00:00 UTC. Panel (a) shows the back-trajectories started at 19 June 2022 12:00 UTC from the box in eastern Germany marked in black. Panel (b) shows the back-trajectories started at 18 June 2022 12:00 UTC from the box in southern France marked in black.

through from above (not shown here). Hence, high near-surface temperatures in the local BL and cooling of the trajectories due to radiation and re-evaporation of precipitation make it possible for AD air to enter the local BL.

3.4.3 Thunderstorms

It has been described for EMLs that the capping inversion due to the warm, well-mixed layer aloft suppresses thunderstorm formation in the centre of the EML, while thunderstorms tend to erupt violently along its edges (Carlson and Ludlam, 1968; Carlson et al., 1983; Farrell and Carlson, 1989; Keyser and Carlson, 1984; Dahl and Fischer, 2016). The analysis of the case presented here suggests that this behaviour is similar for the more general case of an AD. During 18 and 19 June 2022, the majority of the region covered by the AD experiences no lightning, except for a line parallel to the cold front, where lightning occurs during the early hours of 19 June 2022 (Fig. 1, bottom right). This implies that even though the lid is not strong in this case, the warm AD air aloft still suppresses thunderstorm formation in most parts. The thunderstorms that do occur are close to the edge of the 800 hPa AD layer. Only in the early evenings of 16 and 17 June 2022, there is strong lightning activity over the Iberian peninsula (not shown here) which is likely due to typical summer heat thunderstorms which are not suppressed by the

https://doi.org/10.5194/egusphere-2024-2143
Preprint. Discussion started: 17 July 2024

warm AD air aloft, although it has already been there since the early hours of 16 June 2022. During 18 and 19 June 2022, there is also some lightning activity over Spain and the Gulf of Biscay, in the vicinity of the occluding low pressure system (Fig. 1).

Carlson and Ludlam (1968) and Carlson et al. (1983) argue that thunderstorms erupt along the edge of the EML, because heated, moistened air from underneath the lid *underruns* it. It then reaches the edge, where the lid is higher and weaker, so that the constraint is weaker, convection can penetrate and thunderstorms can erupt. Similarly, along the edge, surface heating by insolation can suffice to overcome the lid. Using the Sawyer-Eliasson equation and quasi-geostrophic theory, Keyser and Carlson (1984) conclude that due to the confluence at the midlevel baroclinic zone, a thermally direct circulation develops there, while anti-cyclonic shear induces a cell of thermally indirect circulation in the upper parts of the EML. These circulations act together to induce a branch of rising motion at the midlevel baroclinic zone. They argue, however, that it should be too weak to cause thunderstorm outbreak and rather supports other mechanisms by further weakening the lid. Similarly, Dahl and Fischer (2016), who used Q-vector analysis, found that in the warm season, when EML are influenced by a cyclone's low-level wind field, a convergence line forms along the western edge of the EML, east of the 850 hPa cold front, facilitating lifting and the eruption of thunderstorms. Also Cordeira et al. (2017) argue that the presence of an EML can support strong thunderstorms, if meso–synoptic-scale lifting mechanisms initiate convection.

A closer analysis of the vertical cross-sections in Figs. 4 and 5 can give some insight into the processes involved in the thunderstorm formation in this case. There are two accumulations of lightning along the latitudinal cross-section in Fig. 4. The first is located between 1.25° W and 0.5° E. This coincides with the location of the surface front, at which the air mass above is lifted (red shading in Fig. 4d). In the upper part of the red shaded area (above 4 km), the potential temperature increases slowly with height (Fig 4a), while the equivalent potential temperature decreases with height (Fig. 4b), which indicates potential instability. As the approach of the front lifts the air mass aloft, this potential instability is released and thunderstorms erupt. Relative humidity close to 100 % (Fig. 4c) is also an indicator of the clouds that form in response to the convection.

The second accumulation of lightning is located between 4.75° and 5.75° W. This region is close to the eastern edge of the 800 hPa AD layer (see Figs. 1 and 4) and has a very strong horizontal temperature gradient at the surface, preceding the surface cold front. Here, air converges (not shown here), there is rising motion present associated with a thermally direct circulation, and between 2 and 4 km altitude the equivalent potential temperature decreases. Hence, the lifting can release potential instability. This supports the argument by Keyser and Carlson (1984) and Dahl and Fischer (2016), although there seems to be no thermally indirect circulation that contributes to the rising motion. Around 8° W, in the vicinity of the western edge of the AD at 800 hPa, there is another region with strong ascending motion. Here, this rising branch is associated with a thermally indirect cell. While there was no lightning recorded in this region, ERA5 shows some convective precipitation, which indicates that convection did erupt, although not producing any thunderstorms.

In the longitudinal cross-section (Fig. 5), lightning occurs between 51.5° and 53.25° N, a region with large scale upward motion (red shading above 2 km in Fig. 5d). This region is located north of the surface front, but south of the 800 hPa front and the AD edge at that level. The strongest upward motion is found above 3 km altitude between 53° and 54° N. Also here there is a zone of confluence (not shown here), which might be causing the ascent. Centred at the region with strongest upward motion, there is a wider region where ERA5 produces both large scale and convective precipitation in response to the lifting (Fig. 5e).

https://doi.org/10.5194/egusphere-2024-2143
Preprint. Discussion started: 17 July 2024

Again, dark blue shading above 4 km altitude in Fig. 5c indicates the presence of clouds. In this case, there is increased CAPE at the locations of the thunderstorms, which is released (not shown here). However, the rising motion is not surface-based, which it should be if underrunning were the reason for it.

4 Conclusions

In this study we introduced the concept of atmospheric deserts (ADs), which can be seen as a generalization of elevated mixed layers (EMLs), as they are air masses originating in the dry, hot, convective boundary layers (BLs) of semi-arid, desert, subtropical and/or elevated source regions that are advected to an often cooler, moister target region. Since they progressively lose their distinct characteristics during the advection over hundreds to thousands of kilometers, they cannot be detected based on their thermodynamic properties in the target region.

We introduced a direct detection method, tracing the air mass directly from source to target using Lagrangian trajectories. This allows a detailed and high-resolution analysis of these air masses and their developments during the advection. Different trajectories travel along different paths and also experience different diabatic processes that change their properties. A clustering of the trajectories is used to analyze the typical pathways.

A case study for 15–19 June 2022 is used as an example to explain the involved processes. The AD in this case travels across large parts of southern and central Europe during the duration of the case study and occupies many layers in the vertical between the surface and 13 km, extending further to the surface in its centre, and residing higher aloft at its edges. The 800 hPa cold front approaches the AD from the north west, but only catches up with it by the end of the case study.

Four different trajectory pathways were identified. Of the four, only one cluster behaves like one could expect of a typical EML: It rises slightly over the colder, local air mass, while almost conserving its potential temperature and water vapour mixing ratio and, therefore, the well-mixed properties of the source region's CBL. Diabatic process, however, modify the properties of the trajectories in the other clusters. One cluster rises much higher, thereby cooling adiabatically, which induces condensation. Latent heat causes the potential temperature in this cluster to rise and condensate precipitates out. Another cluster experiences a descent (after an initial ascent together with the other clusters). Meanwhile, its specific water vapour content increases and its potential temperature decreases. This is partly due to radiative cooling, and partly due to re-evaporation of precipitation falling through from above. Also, mixing with the cooler, moister local air can be a reason for the cooling and moistening. A fourth cluster is distinguished mainly by the deviation in its geographical location, but more difficult to interpret due to its heterogeneity and less interesting because it does not travel across central Europe.

ADs can have similar consequences for the weather in the target region as were described for EML: Europe experienced a heat wave during this AD event, and thunderstorm eruption was limited to a narrow line. In this case study, however, the AD did not form a lid for long enough for heat to build up underneath. An analysis of back-trajectories from heat-affected regions indicates that subsidence heating contributes to the increase of near-surface temperatures. When the near-surface temperatures are increased, it is also possible for AD air that is cooled during the advection to penetrate the local BL. This happens for a small percentage of the trajectories that experience strong cooling due to a combination of radiation and evaporation.

https://doi.org/10.5194/egusphere-2024-2143
Preprint. Discussion started: 17 July 2024

Thunderstorms erupted in the vicinity of the occluding low pressure system over the Gulf of Biscay, and along a line, parallel to the edge of the AD and the approaching cold front. Further analysis showed that some of the thunderstorms erupt in the close vicinity of the surface cold front, when it lifts the potentially unstable, overlying air mass. Another region of ascending motion and thunderstorms was found between the surface and the 800 hPa front, close to the edge of the AD at the 800 hPa level. Several processes were suggested to cause the eruption of thunderstorms along the edges of such an air mass. The present case study supported some of the arguments from literature, namely ascending motion due to confluence at the edge, but the data is not sufficient to find more causal relations about why the thunderstorms erupt exactly where they do.

The case study presented in this work helped to understand the processes modifying AD air along the way and the influence of the AD on local weather in the target region. Future work is planned to generalize the results presented here and investigate the processes related to heat waves and thunderstorm formation during ADs in more detail. It will be interesting to study whether the heat buildup under the lid really plays a minor role compared to other mechanisms leading to high near-surface temperatures in the presence of an AD, or whether this was case-specific.

Code availability. The code used to calculate the trajectories and results and plots presented here can be found at: https://doi.org/10.5281/zenodo.12663679

Data availability. ERA5 data is freely available at the Copernicus Climate Change Service (C3S) Climate Data Store (Hersbach et al., 2023). The results contain modified Copernicus Climate Change Service information 2020. Neither the European Commission nor ECMWF is responsible for any use that may be made of the Copernicus information or data it contains. Lightning data from Blitzortung.org is available as participant of measurement network. The LAGRANTO is available from: Sprenger and Wernli (2015).

Author contributions. GM and AZ acquired the funding for this project. FF conducted the calculations, the analysis, and wrote the manuscript under supervision by GM and AZ, with the support of IS and RS. IS acquired the data and RS supported in software development. GM, AZ, IS, and RS reviewed the manuscript prepared by FF.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We thank Deborah Morgenstern for setting up LAGRANTO and running the first explorative trajectory calculations. We thank all colleagues who were involved in discussions. The computational results presented have been achieved [in part] using the Vienna Scientific Cluster (VSC). This work of Fiona Fix was funded by the Austrian Science Fund (FWF, grant no. P35780).

https://doi.org/10.5194/egusphere-2024-2143
Preprint. Discussion started: 17 July 2024

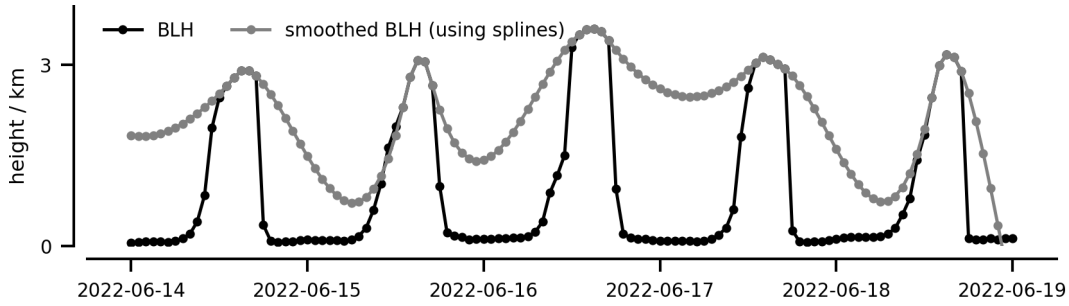

Figure A1. Boundary layer height above mean sea level (black) and smoothed boundary layer height above mean sea level using splines (grey) for one location the source region: 5° E, 30° N.

Appendix A

The result of applying *splines* to the BLH is shown in Fig. A1 for one example location and a time window of 5 days.

Fig. A2 shows cluster C2 (blue) as in Fig. 2, but together with the cluster median, the interquartile range and three individual trajectories from that cluster. From panel (a) it becomes obvious that the individual trajectories experience the jump in altitude at very different times and the shape of the mean and median might that there are two dominant times for this to happen: around 24 h and around 80 h. The changes in θ, T, and q are consistent with this behaviour (panels (b)–(d)). From panels (e) to (f) it becomes apparent, that the mean and median in the cloud water content variables differ considerably. This can be explained by the large spread between the individual trajectories and the skewed distribution. Since each trajectory shows peaks in those variables at slightly different times, there are many trajectories with little cloud water content at all times, and only few with very high values. Panel (c) shows that on average the trajectory temperature only sinks below the freezing level (marked by horizontal line) around hour 60. The fact that there is frozen cloud water content present even before this (panel (e)) is due to the same reasons as the discrepancy in the mean and median. If one pays attention to the three individual trajectories, it can be seen that they do have physically consistent time series in all variables. Hence, it should be kept in mind that the cluster mean (or median) do not represent a physically consistent trajectory. They are, however, still useful to discuss average behaviour of the trajectories in the cluster.

https://doi.org/10.5194/egusphere-2024-2143
Preprint. Discussion started: 17 July 2024

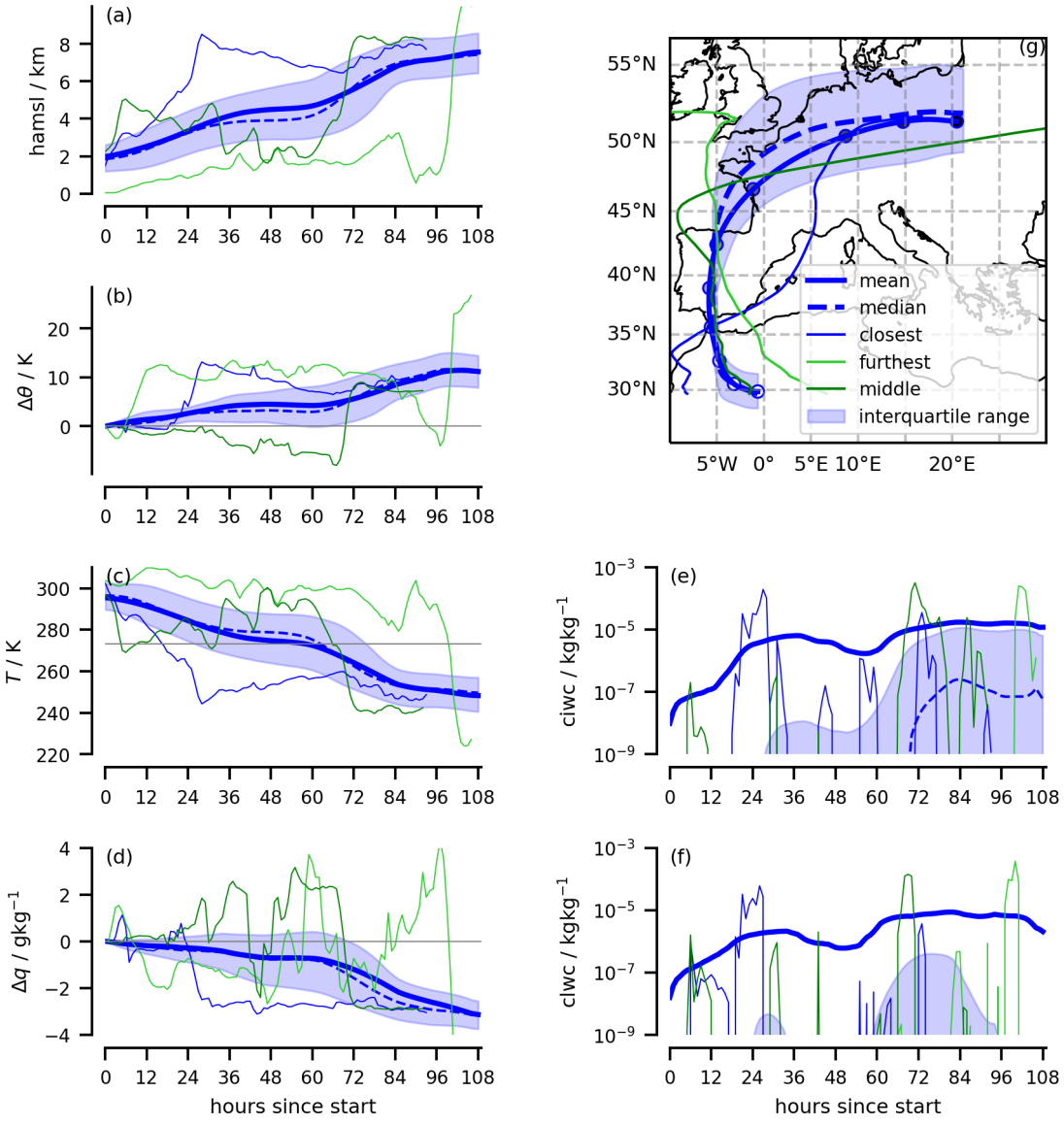

Figure A2. Detailed depiction of the blue cluster (C2) of the trajectories started on 15 June 2022, as in Fig. 2. The thick blue line denotes the cluster mean, the dashed blue line the cluster median, and the shaded region marks the interquartile range. The thin blue line marks the trajectory that is closest to the cluster centroid, the light green one the one that is furthest from the centroid, and the dark green is a trajectory with a medium distance to the cluster centroid. Panel (a): height above mean sea level ($hamsl$). Panel (b): Difference in potential temperature (θ) since initialization. Panel (c) Temperature (T). Panel (d): Difference in specific water content (q) since initialization. Panel (e): cloud ice water content ($ciwc$). Panel (f): cloud liquid water content ($clwc$). Panel (g): Map of the trajectory pathways. Panels (e–f) have a logarithmic scale on the y-axis.

https://doi.org/10.5194/egusphere-2024-2143
Preprint. Discussion started: 17 July 2024

References

Arritt, R. W., Wilczak, J. M., and Young, G. S.: Observations and Numerical Modeling of an Elevated Mixed Layer, Monthly Weather Review, 120, 2869–2880, https://doi.org/10.1175/1520-0493(1992)120<2869:OANMOA>2.0.CO;2, 1992.

Banacos, P. C. and Ekster, M. L.: The association of the elevated mixed layer with significant severe weather events in the northeastern United States, Weather and Forecasting, 25, 1082–1102, https://doi.org/10.1175/2010WAF2222363.1, 2010.

Browning, K. A.: Radar measurements of air motion near fronts, Weather, 26, 320–340, https://doi.org/10.1002/j.1477-8696.1971.tb04211.x, 1971.

Carlson, T. N.: Airflow Through Midlatitude Cyclones and the Comma Cloud Pattern, Monthly Weather Review, 1980.

Carlson, T. N. and Ludlam, F. H.: Conditions for the occurrence of severe local storms, Tellus, 20, 203–226, https://doi.org/10.1111/j.2153-3490.1968.tb00364.x, 1968.

Carlson, T. N., Anthes, R. A., Schwartz, M., Benjamin, S. G., and Baldwin, D. G.: Analysis and Prediction of Severe Storms Environment, 1980.

Carlson, T. N., Benjamin, S. G., and Forbes, G. S.: Elevated Mixed Layers in the Regional Severe Storm Environment: Conceptual Model and Case Studies, Monthly Weather Review, 111, https://doi.org/10.1175/1520-0493(1983)111<1453:EMLITR>2.0.CO;2, 1983.

Cordeira, J. M., Metz, N. D., Howarth, M. E., and Galarneau, T. J.: Multiscale upstream and in situ precursors to the elevated mixed layer and high-impact weather over the Midwest United States, Weather and Forecasting, 32, 905–923, https://doi.org/10.1175/WAF-D-16-0122.1, 2017.

Dahl, J. M. and Fischer, J.: The origin of western European warm-season prefrontal convergence lines, Weather and Forecasting, 31, 1417–1431, https://doi.org/10.1175/WAF-D-15-0161.1, 2016.

European Centre for Medium-Range Weather Forecasts: IFS Documentation CY41R2 - Part IV: Physical Processes, vol. 4, European Centre for Medium-Range Weather Forecasts, https://doi.org/10.21957/tr5rv27xu, 2016.

European Commission, Joint Research Centre, Hill, J., Von Maltitz, G., Sommer, S., Reynolds, J., Hutchinson, C., and Cherlet, M.: World atlas of desertification – Rethinking land degradation and sustainable land management, Publications Office, https://doi.org/doi/10.2760/06292, 2018.

Farrell, R. J. and Carlson, T. N.: Evidence for the Role of the Lid and Underrunning in an Outbreak of Tornadic Thunderstorms, Monthly Weather Review, 117, 857–871, https://doi.org/10.1175/1520-0493(1989)117<0857:EFTROT>2.0.CO;2, 1989.

Garcia-Carreras, L., Parker, D. J., Marsham, J. H., Rosenberg, P. D., Brooks, I. M., Lock, A. P., Marenco, F., Mcquaid, J. B., and Hobby, M.: The turbulent structure and diurnal growth of the Saharan atmospheric boundary layer, Journal of the Atmospheric Sciences, 72, 693–713, https://doi.org/10.1175/JAS-D-13-0384.1, 2015.

Gitro, C. M., Bikos, D., Szoke, E. J., Jurewicz, M. L., Cohen, A. E., and Foster, M. W.: A Demonstration of Modern Geostationary and Polar-Orbiting Satellite Products for the Identification and Tracking of Elevated Mixed Layers, Journal of Operational Meteorology, 7, 180–192, https://doi.org/10.15191/NWAJOM.2019.0713, 2019.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., Chiara, G. D., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Vil-

https://doi.org/10.5194/egusphere-2024-2143
Preprint. Discussion started: 17 July 2024

laume, S., and Thépaut, J. N.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Sabater, J. M., Nicolas, J., Peubey, C., R., R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., and Thépaut, J.-N.: ERA5 hourly data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), accessed on 18-03-2024, 2023.

Imbery, F., Friedrich, K., Fleckenstein, R., Kaspar, F., Bissolli, P., Haeseler, S., Daßler, J., and Kreis, A.: Intensive Hitzewelle im Juni 2022 in Deutschland und Mitteleuropa, https://www.dwd.de/DE/leistungen/besondereereignisse/temperatur/20220629_temperatur_hitzewelle-juni.pdf?__blob=publicationFile&v=5, 2022.

Johns, R. H. and Dorr, R. A.: Some meteorological aspects of strong and violent tornado episodes in New England and Eastern New York, National Weather Digest, 20, 2–12, 1996.

Karyampudi, V. M. and Carlson, T. N.: Analysis and Numerical Simulations of the Saharan Air Layer and Its Effect on Easterly Wave Disturbances, Journal of the Atmospheric Sciences, 45, 3102–3136, 1988.

Keune, J., Schumacher, D. L., and Miralles, D. G.: A unified framework to estimate the origins of atmospheric moisture and heat using Lagrangian models, Geosci. Model Dev, 15, 1875–1898, https://doi.org/10.5194/gmd-15-1875-2022, 2022.

Keyser, D. and Carlson, T. N.: Transverse Ageostrophic Circulations Associated with Elevated Mixed Layers, Monthly Weather Review, https://doi.org/10.1175/1520-0493(1984)112<2465:TACAWE>2.0.CO;2, 1984.

Lanicci, J. M. and Warner, T. T.: A synoptic climatology of the elevated mixed-layer inversion over the southern Great Plains in spring. Part I: Structure, dynamics, and seasonal evolution, Weather and Forecasting, 6, 181–197, https://doi.org/10.1175/1520-0434(1991)006<0181:ASCOTE>2.0.CO;2, 1991a.

Lanicci, J. M. and Warner, T. T.: A synoptic climatology of the elevated mixed-layer inversion over the southern Great Plains in spring. Part II: The life cycle of the lid, Weather and Forecasting, 6, 198–213, https://doi.org/10.1175/1520-0434(1991)006<0198:ASCOTE>2.0.CO;2, 1991b.

Lanicci, J. M. and Warner, T. T.: A synoptic climatology of the elevated mixed-layer inversion over the southern Great Plains in spring. Part III: Relationship to severe-storms climatology, Weather and Forecasting, 6, 214–226, https://doi.org/10.1175/1520-0434(1991)006<0214:ASCOTE>2.0.CO;2, 1991c.

Lewis, M. W. and Gray, S. L.: Categorisation of synoptic environments associated with mesoscale convective systems over the UK, Atmospheric Research, 97, 194–213, https://doi.org/10.1016/J.ATMOSRES.2010.04.001, 2010.

MacQueen, J.: Some Methods for Classification and Analysis of MultiVariate Observations, https://www.bibsonomy.org/bibtex/25dcdb8cd9fba78e0e791af619d61d66d/enitsirhc, 1967.

Nie, Y. and Sun, J.: Moisture sources and transport for extreme precipitation over Henan in July 2021, Geophysical Research Letters, 49, e2021GL097 446, https://doi.org/10.1029/2021GL097446, 2022.

Oertel, A., Pickl, M., Quinting, J. F., Hauser, S., Wandel, J., Magnusson, L., Balmaseda, M., Vitart, F., and Grams, C. M.: Everything hits at once: How remote rainfall matters for the prediction of the 2021 North American heat wave, Geophysical Research Letters, 50, e2022GL100 958, https://doi.org/10.1029/2022GL100958, 2023.

Ribeiro, B. Z. and Bosart, L. F.: Elevated mixed layers and associated severe thunderstorm environments in South and North America, Monthly Weather Review, 146, 3–28, https://doi.org/10.1175/MWR-D-17-0121.1, 2018.

Schär, C.: Climate extremes: The worst heat waves to come, Nature Climate Change, 6, 128–129, https://doi.org/10.1038/nclimate2864, 2016.

https://doi.org/10.5194/egusphere-2024-2143
Preprint. Discussion started: 17 July 2024

Schär, C. and Jendritzky, G.: Hot news from summer 2003, Nature, 432, 559–560, https://doi.org/https://doi.org/10.1038/432559a, 2004.

Sibley, A.: Thunderstorms from a Spanish plume event on 28 June 2011, Weather, 67, 143–146, https://doi.org/10.1002/wea.1928, 2012.

Sprenger, M. and Wernli, H.: The LAGRANTO Lagrangian analysis tool - Version 2.0, Geoscientific Model Development, 8, 2569–2586, https://doi.org/10.5194/gmd-8-2569-2015, 2015.

Stohl, A., Haimberger, L., Scheele, M. P., and Wernli, H.: An intercomparison of results from three trajectory models, Meteorological Applications, 8, 127–135, https://doi.org/10.1017/S1350482701002018, 2001.

Tchakerian, V. P.: HYDROLOGY, FLOODS AND DROUGHTS | Deserts and Desertification, Encyclopedia of Atmospheric Sciences: Second Edition, pp. 185–192, https://doi.org/10.1016/B978-0-12-382225-3.00035-9, 2015.

van der Does, M., Knippertz, P., Zschenderlein, P., Harrison, R. G., and Stuut, J. B. W.: The mysterious long-range transport of giant mineral dust particles, Science Advances, 4, https://doi.org/10.1126/sciadv.aau2768, 2018.

Vaughn, D. M.: Arid climates, Encyclopedia of Earth Sciences Series, pp. 85–89, https://doi.org/10.1007/1-4020-3266-8_16/TABLES/1, 2005.

Wanke, E., Andersen, R., and Volgnandt, T.: A world-wide low-cost community-based time-of-arrival lightning detection and lightning location network (Project description; 11 May 2014), https://www.blitzortung.org/en/cover_your_area.php, 2014.

Acknowledgments

I would like to express my gratitude to the many individuals who contributed to the completion of this habilitation. Their support, expertise, and encouragement have been invaluable over the past years.

Firstly, I am deeply thankful to my academic advisors, especially—but not limited to—Achim Zeileis (Department of Statistics) and Georg J. Mayr (Department of Atmospheric and Cryospheric Sciences). Their support and belief in my potential have been a constant source of motivation, even during times when the seas were rougher, the batteries low, and the deadlines tighter than usual.

I also extend my sincere thanks to all my colleagues and the wonderful team I have the pleasure of working with, both at the Faculty of Economics and Statistics and at the Digital Science Center.

On a personal note, I am deeply appreciative of the support and understanding of my family and friends, with whom I could share not only the best moments but also the challenging ones. They have always been a source of motivation, offering unwavering support whenever needed.

Thank you all for your invaluable support!

Additional References

Breiman, L., 2001: Random forests. *Machine Learning*, **45 (1)**, 5–32, doi:10.1023/A:1010933404324.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone, 1984: Classification and regression trees (1st edition). 1–368, doi:10.1201/9781315139470.

Buizza, R., and R. David, 2017: 25 years of ensemble forecasting at ECMWF. doi:10.21957/bv418o.

Dabernig, M., G. J. Mayr, J. W. Messner, and A. Zeileis, 2017: Spatial ensemble post-processing with standardized anomalies. *Quarterly Journal of the Royal Meteorological Society*, **143 (703)**, 909–916, doi:doi.org/10.1002/qj.2975.

Dusch, M., 2019: foehnix: A toolbox for automated foehn classification based on mixture models. URL https://matthiasdusch.github.io/foehnix-python/, python package version 0.1.2.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *Journal of Applied Meteorology and Climatology*, **11 (8)**, 1203–1211, doi:10.1175/1520-0450(1972)011<1203:TUOMOS>2.0.CO;2.

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, **133 (5)**, 1098–1118, doi:10.1175/MWR2904.1.

Hawkins, E., D. McNeall, D. Stephenson, J. Williams, and D. Carlson, 2014: The end of the rainbow: An open letter to the climate science community. URL https://www.climate-lab-book.ac.uk/2014/end-of-the-rainbow/, accessed 2024-11-11.

Hothorn, T., and A. Zeileis, 2015: partykit: A modular toolkit for recursive partytioning in *R*. *Journal of Machine Learning Research*, **16 (118)**, 3905–3909, URL http://jmlr.org/papers/v16/hothorn15a.html.

Hunter, J., D. Dale, E. Firing, M. Droettboom, and Matplotlib development team, 2017: *What's new in Matplotlib 2.0: Changes to the default style*. URL https://matplotlib.org/stable/users/prev_whats_new/dflt_style_changes.html, accessed 2024-11-11.

Klein, N., T. Kneib, S. Lang, and A. Sohn, 2015: Bayesian structured additive distributional regression with an application to regional income inequality in Germany. *The Annals of Applied Statistics*, **9 (2)**, 1024–1052, doi:10.1214/15-AOAS823.

Lang, M. N., L. Schlosser, and A. Zeileis, 2024: *circtree: Regression Trees and Forests for Circular Responses*. *R* package version 0.1-0.

Messner, J. W., G. J. Mayr, and A. Zeileis, 2016: Heteroscedastic censored and truncated regression with crch. *The R Journal*, **8 (1)**, 173–181, doi:10.32614/RJ-2016-012.

Rigby, R. A., and D. M. Stasinopoulos, 2005: Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54 (3)**, 507–554, doi:https://doi.org/10.1111/j.1467-9876.2005.00510.x.

Schefzik, R., T. L. Thorarinsdottir, and T. Gneiting, 2013: Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science*, **28 (4)**, 616–640, doi:10.1214/13-STS443.

Schlosser, L., M. N. Lang, T. Hothorn, and A. Zeileis, 2023: *disttree: Trees and Forests for Distributional Regression*. *R* package version 0.2-1.

Stauffer, R., 2023: foehnix: A toolbox for automated foehn classification based on mixture models. URL https://retostauffer.github.io/Rfoehnix/, *R* package version 0.1.6.

Umlauf, N., N. Klein, T. Simon, and A. Zeileis, 2021: bamlss: A Lego toolbox for flexible Bayesian regression (and beyond). *Journal of Statistical Software*, **100 (4)**, 1–53, doi:10.18637/jss.v100.i04.

Umlauf, N., N. Klein, and A. Zeileis, 2018: BAMLSS: Bayesian additive models for location, scale, and shape (and beyond). *Journal of Computational and Graphical Statistics*, **27 (3)**, 612–627, doi:10.1080/10618600.2017.1407325.

Wickham, H., 2016: *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, URL https://ggplot2.tidyverse.org.

Zeileis, A., and P. Murrell, 2023: Coloring in R's blind spot. *The R Journal*, **15**, 240–256, doi:10.32614/RJ-2023-071.

Zhu, Y., and R. E. Newell, 1998: A proposed algorithm for moisture fluxes from atmospheric rivers. *Monthly Weather Review*, **126 (3)**, 725–735, doi:10.1175/1520-0493(1998)126<0725:APAFMF>2.0.CO;2.