Statistical Postprocessing of Numerical Weather Prediction Forecasts using Machine Learning

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

von der KIT-Fakultät für Mathematik des
Karlsruher Instituts für Technologie (KIT)
genehmigte

DISSERTATION

von

Benedikt Schulz, M.Sc.
geboren in Kandel

Tag der mündlichen Prüfung: 17. Mai 2023

Referent: Prof. Dr. Tilmann Gneiting
Korreferenten: Prof. Dr. Peter Knippertz
Dr. Sebastian Lerch

Acknowledgments

My deepest gratitude belongs to Sebastian Lerch, Peter Knippertz and Tilmann Gneiting for their excellent supervision over the last years. Through their guidance, support, experience and example, I was encouraged and able to pursue my doctoral studies, during which I have greatly enjoyed the interdisciplinary working environment being affiliated to research groups at three different faculties.

The research presented in this thesis has been funded by the Deutsche Forschungsgemeinschaft (DFG) through the subproject "C5 – Dynamical feature-based ensemble postprocessing of wind gusts within European winter storms" of the Transregional Collaborative Research Center SFB / TRR 165 "Waves to Weather" and is gratefully acknowledged.

Over the course of my doctoral studies, I have benefited from the help and advice of many colleagues. First, I would like to thank Lea Eisenstein for being a great project partner in C5, whom I could always ask for advice and have fun with. Also, I would like to thank Eva-Maria Walz, Alexander Jordan, Johannes Resin, Ghulam Qadir, Daniel Wolffram, and all other members of the Computational Statistics group; Michael Maier-Gerber for an educational interdisciplinary collaboration, Marco Wurth for assistance on the KIT-Weather portal, Andreas Fink, and all other members of the working group Atmospheric Dynamics; Nina Horat, Jieyu Chen, and all other members of the Young Investigator Group "Artificial Intelligence for Probabilistic Weather Forecasting"; Tamara Göll for a semester full of thrilling teaching activities, Steffen Betsch, and all other members of the Institute for Stochastics; and Kevin Höhlein, and all early career scientists and members of Waves to Weather for many stimulating discussions and meetings.

Finally, I am grateful and feel very fortunate in light of the great support of my family and friends.

Contents

CHAPTER 1

Introduction

In the face of an uncertain future, humans have sought to shed light on forthcoming events for a long time. However, "predictions are difficult, especially about the future".[1] A key message underlying this famous saying is that the future is inherently uncertain, and that forecasts ought to be probabilistic in nature. Scientific theories that challenged the prevalent believe of a fully predictable, deterministic universe, such as chaos theory or quantum mechanics, have stimulated the use of probabilistic forecasts that take the form of full predictive distributions. Nowadays, quantifying forecast uncertainty provides valuable information for decision making in fields such as economics, demography, or hydrology, among others.

Arguably, the most influential application of probabilistic forecasting is in weather prediction. Based on the chaos theory (Lorenz, 1963), meteorologists found that the predictability of the atmosphere is limited and identified the need to quantify the forecast uncertainty. Due to the fact that making and evaluating forecasts goes hand in hand, meteorologists contributed massively to research on statistical forecasting. Nowadays, weather prediction is based on numerical models of the physics of the atmosphere that generate probabilistic forecasts, which represent distinct scenarios of the future. Next to agricultural, economic, recreational and transportation activities, weather forecasts (e.g., of wind and solar irradiance) are becoming even more important in light of the shift towards renewable energy sources. However, probabilistic forecasts from numerical weather prediction models are subject to systematic errors such as biases and dispersion errors. In a nutshell, they lack accuracy and do not quantify the forecast uncertainty adequately. Based on past forecast-observation pairs, statistical methods can learn to correct the numerical forecasts for these errors, in order to generate accurate and reliable probabilistic forecasts. We refer to this process as statistical postprocessing.

In the last decade, methods from machine learning have seen unprecedented rise and success

[1] This Danish proverb has appeared in several variants and is often attributed to Niels Bohr, although emerging earlier (`https://quoteinvestigator.com/2013/10/20/no-predict/`).

in research and application. As a result, artificial intelligence is inevitably becoming an indispensable aspect of daily life and receiving more and more attention in public, e.g., due to artificial intelligence models such as ChatGPT.[2] Artificial neural networks have proven to be powerful methods for forecasting, and are therefore a promising pathway for progress in probabilistic weather prediction. First applications of modern machine learning methods in the context of probabilistic weather prediction have demonstrated huge potential benefits (e.g., Rasp and Lerch, 2018; Haupt et al., 2021).

This thesis investigates the use of modern machine learning methods in the context of statistical postprocessing. Bridging a gap between the classic literature on statistical forecasting and artificial neural networks, we specifically highlight aspects of aggregating distributional forecasts of neural network ensembles. In doing so, we aim to develop concrete recommendations for the aggregation of probabilistic forecasts generated by the ensembles, finding the best method for combination and the optimal ensemble size. Due to their potential hazardousness in winter storms and relevance for energy generation, wind gusts are of special societal importance. Still, research on statistical postprocessing of forecasts of wind gusts is scarce. In this thesis, we present a wide range of statistical postprocessing methods for wind gusts based on modern machine learning, with a focus on European winter storms. In comprehensive case studies, we assess the predictive performance of the methods in order to identify the best postprocessing method for wind gusts. Further, we investigate the additional value of using machine learning for postprocessing. At last, we address the question of incorporating domain knowledge into machine learning models, at the example of European winter storms. To that end, we develop an objective identification of specific meteorological conditions, and demonstrate how to evaluate what the model has learned, addressing the research question of interpretable machine learning.

Statistical postprocessing exemplifies the duality of forecast generation and evaluation, as it is motivated by systematic errors of numerical weather predictions. In order to correct these errors, we first need to know how goodness in terms of distributional forecasts is characterized. Chapter 2 introduces theoretical background for statistical forecasting, which builds the foundation for the development of probabilistic forecasting methods. We introduce the concepts of calibration and sharpness in order to formulate the central paradigm of probabilistic forecasting. A predictive distribution can take different forms, we present three exemplary types that are of particular importance throughout the work. Following the introductory chapters, the thesis begins with Chapter 3, where we investigate a problem that arises in the development of neural network-based postprocessing methods, namely, the aggregation of distributional forecasts from ensembles of neural networks (so-called deep ensembles). Since the scope of this problem is not bound to the application in statistical postprocessing, we cover it in a general setting. As combining predictive distributions is a well-known problem in the field of statistical forecasting, and deep ensembles have been proven

[2] `https://openai.com/blog/chatgpt`

to generate state-of-the-art forecasts, the contribution of Chapter 3 is to perform a systematic analysis of the aggregation of distributional forecasts from deep ensembles.

The main part of the thesis begins with Chapter 4, which illustrates the need for statistical postprocessing and presents a wide range of techniques thereof. The complexity of the methods introduced ranges from basic approaches rooted in classical statistics to complex machine learning methods. We focus on the development of a neural network-based framework for postprocessing building on ideas from Chapter 2 and the results of Chapter 3. In Chapter 5, we apply the statistical postprocessing methods presented before and assess their predictive performance with respect to the central concepts of calibration and sharpness. Three case studies that highlight different aspects of statistical postprocessing are presented. An emphasis lies on the third case study concerned with wind gust prediction, for which we perform a systematic comparison of the methods presented before.

Generic machine learning methods are often able to achieve high-level predictive performance without the integration of domain knowledge. However, in Chapter 6, we investigate the idea of hybrid weather prediction models that integrate specific meteorological expertise to overcome shortcomings of the approaches presented in the former, demonstrated through the example of European winter storms and tropical cyclones in the North Atlantic. Chapter 7 concludes the thesis with a summary of the findings and potential directions for future research.

1.1 Relation to previous and published work

Large parts of the work presented in this dissertation have resulted in publications over the course of my doctoral studies. These are listed here in order of publication. Further, I state where the publications are integrated in the thesis, and describe the contributions of myself and the coauthors. Table 1.1 provides an overview of the integration of the publications. As the publications cover related topics based on the same principles, the results have been merged, rearranged or extended for a consistent thesis. Hence, direct quotes of these publications appear throughout the thesis without being explicitly marked as such to ensure a better readability. For all publications, all authors contributed to discussions and text revisions. For code, we refer to the publications.

Note that two publications developed from collaborations with meteorologists, where I mostly contributed to the statistical parts of the paper (Maier-Gerber et al., 2021; Eisenstein et al., 2022). To provide context, meteorological parts that I did not substantially contribute to are still included in this thesis. For those publications, we will describe my contributions and those of my meteorological collaborators in more detail.

> **Schulz et al. (2021):** B. Schulz, M. El Ayari, S. Lerch, and S. Baran (2021). "Post-processing numerical weather prediction ensembles for probabilistic solar irradiance forecasting". *Solar Energy*, 220, 1016–1031.

Table 1.1: Overview of the publications underlying this thesis.

Thesis	Publication	Part
Sct. 2.1	Schulz and Lerch (2022b)	App. A
Sct. 2.2	Schulz and Lerch (2022b)	App. A
	Schulz and Lerch (2022a)	Sct. 2.1
Sct. 2.3	Schulz et al. (2021)	Sct. 3.1
	Schulz and Lerch (2022b)	Sct. 3c
	Schulz and Lerch (2022a)	Sct. 3
Ch. 3	Schulz and Lerch (2022a)	All
Ch. 4	Schulz and Lerch (2022b)	Sct. 3
Sct. 4.3	Schulz and Lerch (2022a)	Sct. 3
Sct. 5.1.1	Schulz et al. (2021)	Scts. 2.2 and 3.3
Sct. 5.2	Schulz et al. (2021)	All
Sct. 5.3	Schulz and Lerch (2022b)	Scts. 2–6, App. B
Sct. 5.4	Schulz et al. (2021)	Sct. 6
	Schulz and Lerch (2022b)	Sct. 6
Sct. 6.2	Maier-Gerber et al. (2021)	Scts. 1–3 and 5–7
Sct. 6.3	Eisenstein et al. (2022)	Scts. 1, 3, 4, 6, 7.4 and 8, Apps. B–D

The publication was written jointly by all authors. I and Sebastian Lerch designed the project and wrote the original manuscript based on the application in the case study over Germany. Mehrez El Ayari and Sándor Baran integrated the case study over Hungary, developed the advanced models and conducted the final case studies.

The publication is the basis of Section 5.2, parts of the description of the data in Section 5.1.1 and parts of the discussion in Section 5.4. Furthermore, the description of the censored logistic distribution in Section 2.3.1 is taken from the paper.

> **Maier-Gerber et al. (2021):** M. Maier-Gerber, A. H. Fink, M. Riemer, E. Schoemer, C. Fischer, and B. Schulz (2021). "Statistical-dynamical forecasting of sub-seasonal North Atlantic tropical cyclone occurrence". *Weather and Forecasting*, 36 (6), 2127–2142.

Michael Maier-Gerber, the first author, developed the methods, conducted the case study and wrote the paper. I developed the calibration method for the dynamical model forecast (Section 3b of the paper) and gave advice on the development of the statistical models (Section 5 of the paper) and the forecast evaluation (Sections 2d and 6 of the paper).

The publication is the basis of Section 6.2, where we put an emphasis on the statistical models and the forecast evaluation. The section includes parts of the publication that I have not substantially contributed to, but are necessary to provide context. These are the description of the setting (Section 1 of the paper), of the data (Sections 2a–c of the paper), of

the climatological and unprocessed dynamical models (Sections 3a and b of the paper) and concluding remarks in the context of tropical cyclone forecasting (Section 7 of the paper).

> **Schulz and Lerch (2022b):** B. Schulz and S. Lerch (2022b). "Machine learning methods for postprocessing ensemble forecasts of wind gusts: A systematic comparison". *Monthly Weather Review*, 150 (1), 235–257.

I developed the postprocessing methods, conducted the systematic comparison and wrote the paper under the supervision of Sebastian Lerch.

The publication is the basis for main parts of the thesis, specifically, Chapters 2 and 4, Section 5.3 and parts of the discussion in Section 5.4.

> **Schulz and Lerch (2022a):** B. Schulz and S. Lerch (2022a). "Aggregating distribution forecasts from deep ensembles". *Journal of Machine Learning Research*, under revision. Preprint available at `https://arxiv.org/abs/2204.02291v1`.

Based on the problem of aggregating deep ensembles in Schulz and Lerch (2022b), I conducted the simulation and case study, and wrote the paper under the supervision of Sebastian Lerch.

The preprint is the basis of Chapter 3 that is in large parts taken from the paper. The description of the neural network-based postprocessing methods coincides with that in Schulz and Lerch (2022b) and was also used to write Section 2.3 and 4.3. In addition, small parts of Chapter 2 are taken from the paper.

> **Eisenstein et al. (2022):** L. Eisenstein, B. Schulz, G. A. Qadir, J. G. Pinto, and P. Knippertz (2022). "Identification of high-wind features within extratropical cyclones using a probabilistic random forest – Part 1: Method and case studies". *Weather and Climate Dynamics*, 3 (4), 1157–1182.

The first author Lea Eisenstein wrote the paper and conducted the meteorological part of the analysis, while I contributed the statistical part, specifically, the forecast evaluation (Section 3.4, Appendices B1 and C1 of the paper), the development and implementation of the random forest method (Section 4.2, Appendix B2 of the paper), the application of the Kriging method (Section 4.3, Appendix B3 of the paper) and the statistical evaluation (Sections 6.1 and large parts of 6.2 of the paper). The (specific) Kriging method itself was developed and implemented by Ghulam Qadir (Section 4.3, Appendix C2 of the paper).

The publication is the basis of Section 6.3, where an emphasis is put on the statistical aspects of the study. This section includes parts of the publication that I have not substantially contributed to, but are necessary to provide context. These are the description of the setting (Section 1 of the paper), of the data (Sections 3.1–3 of the paper), the subjective labeling of the data (Section 4.1 of the paper) and the meteorological interpretation of the results (parts of Sections 6, 7.4 and 8 of the paper).

CHAPTER 2

Prelude: Theory on Statistical Forecasting

The prediction of the future concerns inherently uncertain events, hence forecasts ought to be probabilistic. Probabilistic forecasts aim to quantify the uncertainty associated with predicting a future quantity or event, while deterministic point forecasts consist only of one number.

In this section, we introduce the mathematical foundation for the theory of statistical forecasting, in particular probabilistic forecasting, as reviewed by Gneiting and Katzfuss (2014). First, we present the general framework for statistical forecasting including the central notions of calibration, dispersion and sharpness following Gneiting and Ranjan (2013). Then, we review techniques used to properly assess predictive performance. At last, we introduce exemplary types of forecast distributions that are of importance throughout this thesis.

2.1 Prediction spaces, calibration and sharpness

In order to study forecasts and their behavior, we are interested in the joint distribution of forecasts and observations. Based on the seminal work of Murphy and Winkler (1987), the joint distribution is modeled via a probability space $(\Omega, \mathcal{A}, \mathbb{Q})$ tailored to the task at hand, which is referred to as *prediction space*. Restricting our attention to the case of real-valued observations, each sample of the probability space is identified with a tuple of the form

$$(F_1, \ldots, F_K, Y)\,, \tag{2.1}$$

where $F_1, \ldots, F_K$, $K \geq 1$, are probability measures on the real line, which we identify with their associated cumulative distribution function (CDF) F. The probability measures $F_1, \ldots, F_K$ are referred to as *probabilistic forecasts* and come from K distinct sources or forecasters such as (statistical) models or experts. Hence, the probability measure $\mathbb{Q}$ on $(\Omega, \mathcal{A})$ models the distribution of the probabilistic forecasts and the observation. In the following, all statements will refer to prediction space setting and the joint distribution $\mathbb{Q}$.

The critical notions of calibration and dispersion are based on the *probability integral transform* (PIT). For continuous CDFs F, the PIT is given by $Z_F = F(Y)$, that is, the value the CDF attains at the observation. In case of discontinuities, it is suitably adapted via the *randomized PIT* $Z_F = F(Y-) + V(F(Y) - F(Y-))$, where $F(y-) := \lim_{x \uparrow y} F(x)$ is the left-hand limit of the CDF F at $y \in \mathbb{R}$ and V is a standard uniformly distributed random variable that is independent of the forecasts and the observation. A well-known property of the PIT is that if $Y \sim F$, the PIT is standard uniform, that is, $Z_F \sim \mathcal{U}(0, 1)$. This property also motivates the first and most critical notion of calibration. Namely, a probabilistic forecast F is *probabilistically calibrated* if its PIT Z_F is standard uniformly distributed. Closely connected to probabilistic calibration are three kinds of dispersion that are used to draw conclusions on the forecast behavior based on the distribution of the PIT. A forecast F with PIT Z_F is *overdispersed* if $\mathrm{Var}(Z_F) < 1/12$, *neutrally dispersed* if $\mathrm{Var}(Z_F) = 1/12$ and *underdispersed* if $\mathrm{Var}(Z_F) > 1/12$. As the variance of a standard uniform distribution is given by $1/12$, it is straightforward to see that probabilistic calibration implies neutral dispersion. In general, the opposite does however not hold. Further, a forecast that is over- or underdispersed is thus necessarily not probabilistically calibrated. In practice, the notions of dispersion are often used interchangeably with that of confidence, where an overdispersed forecast corresponds to an underconfident forecast and vice versa.

PIT histograms are used to qualitatively assess calibration and types of miscalibration. Given a set of forecast-observation pairs $\{(F_1, y_1), \ldots, (F_n, y_n)\}$, we can calculate the PIT values z_j, e.g., via $z_j = F_j(y_j)$ for a continuous CDF F_j, $j = 1, \ldots, n$, and generate a histogram of the observed PIT values. A flat histogram indicates a standard uniform distribution of the PIT values and therefore probabilistic calibration (and neutral dispersion), hence we refer to such a forecast as well-calibrated. In contrast, a U-shaped histogram corresponds to underdispersion, a hump-shaped histogram to overdispersion and a skewed histogram to a bias in the forecast. Figure 2.1 illustrates PIT histograms that exhibit those kinds of miscalibration.

For probabilistic forecasts that are issued in form of a sample $\{x_1, \ldots, x_m\}$ with fixed size m, the *rank* r of the observation y, that is, $r = \#\{x_i \leq y, i = 1, \ldots, m\}$, is calculated instead of the PIT. Under the assumption of calibration, that is, the observation is indistinguishable from the sample, each rank is equally likely and the ranks are thus uniformly distributed on $\{1, \ldots, m + 1\}$. As for the PIT, a histogram of the ranks, referred to as *verification rank histogram* or *Talagrand diagram* can be used to check for calibration and is interpreted analogously to a PIT histogram. If the forecast sample size m is not fixed, the calculation of a rank histogram is not feasible. Instead, the *unified PIT*, a generalized version of the PIT, is used to transform ranks to PIT values via $(r - 1)/(m + 1) + U/(m + 1)$, where U is standard uniform (Vogel et al., 2018). In this work, we calculate PIT values for continuous distributions, randomized PIT values for mixed continuous-discrete distributions (such as the censored logistic distribution introduced in Section 2.3.1) and unified PIT values for forecasts

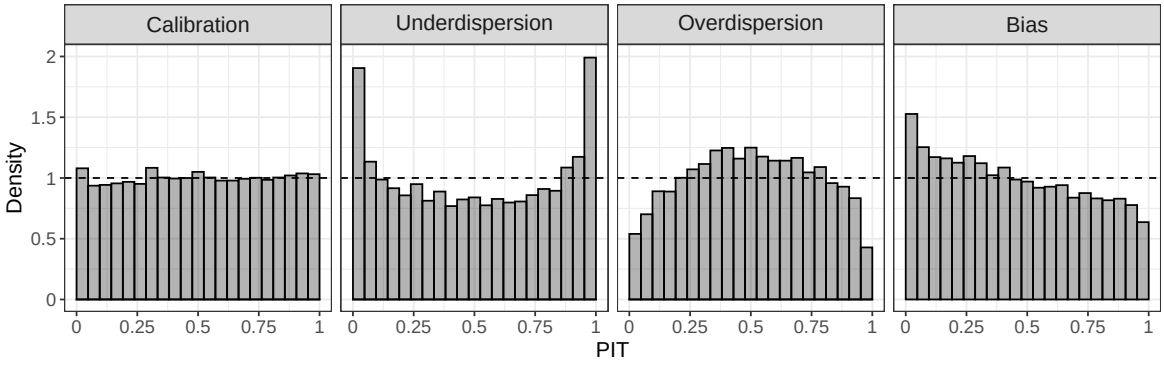

Figure 2.1: Typical shapes of PIT histograms. From left to right, the histograms indicate probabilistic calibration, underdispersion, overdispersion and a bias. The histograms are based on 10,000 observations of a standard normal distribution, which is also the calibrated forecast. The underdispersed forecast is instead based on a standard deviation of $\sigma = 0.8$, the overdispersed on $\sigma = 1.2$ and the biased on a mean of $\mu = 0.2$.

based on samples (of different sizes). As it should be apparent from the type of forecast distribution what variant of the PIT is used, we will not distinguish between the different variants of the PIT.

Prediction intervals (PIs) $[a, b]$ at the $(1-\alpha)$-level provide another type of probabilistic forecast, which is calibrated when $\mathbb{Q}(a \leq Y \leq b) = 1 - \alpha$ holds. In this case, calibration can be assessed quantitatively based on the empirical coverage, that is, the ratio of observations falling in the PI. Given an interval at the $(1-\alpha)$-level, we expect the same ratio of observations to fall within the PI under the assumption of calibration. Given a predictive distribution, we derive a central $(1-\alpha)$ PI based on the quantiles at levels $\alpha/2$ and $1-\alpha/2$. Given a sample $\{x_1, \ldots, x_m\}$ of size $m \in \mathbb{N}$, we can calculate a PI at the $(m-1)/(m+1)$-level via $a = x_{(1)}$ and $b = x_{(m)}$ based on the order statistics $\{x_{(1)}, \ldots, x_{(m)}\}$, e.g., for a sample of size 20, we obtain a PI approximately at the 90.48% level.[1]

In order to state the central paradigm of probabilistic forecasting, we need to introduce the term *sharpness*. A probabilistic forecast is said to be the sharper, the more concentrated, or confident, it is. In probabilistic terms, this refers to the concentration or the spread of the forecast distribution. In practice, sharpness is typically measured in terms of the length of PIs or other quantities that measure the spread of the distribution such as the standard deviation, where in both cases smaller values indicate a sharper forecast. While calibration is a joint property of the forecast and the observation, sharpness is a property of the forecast alone.

The question whether a sharper forecast should be preferred brings us to the central paradigm of probabilistic forecasting, that is, to *maximize the sharpness of the predictive*

[1] In general, $a = x_{(i)}$ and $b = x_{(m+1-i)}$ is a central PI at the $(m+1-2i)/(m+1)$-level with $i = 1, \ldots, \lfloor m/2 \rfloor$.

distribution subject to calibration (Gneiting et al., 2007). Thus, it is only preferable when it is reasonably well-calibrated.

Before investigating the quantitative assessment of probabilistic forecasts, we briefly focus on a popular, special case of probabilistic forecasts. Given a dichotomous target variable $Y \in \{0, 1\}$, a probabilistic forecast takes the form of a Bernoulli distribution and can be identified by the associated success probability p. Therefore, we refer to *probability forecasts* in case of a binary target variable. Calibration can be reduced to one criterion, namely, a probability forecast p for a binary variable Y is called *(conditionally) calibrated* if

$$\mathbb{Q}\left(Y = 1 \mid p\right) = p \quad \text{almost surely.} \tag{2.2}$$

Gneiting and Ranjan (2013, Theorem 2.11) prove that conditional calibration is equivalent to probabilistic calibration and other notions of calibration not presented in this work. Hence, it is sufficient to consider only the calibration criterion of equation (2.2) in practice.

The calibration of probability forecasts is checked via *reliability diagrams*, which allow a qualitative assessment of the calibration criterion (Sanders, 1963). Given a set of forecast-observation pairs $\{(p_1, y_1), \ldots, (p_n, y_n)\}$, a partition of the unit interval is used to assign each pair to a bin. Within each bin, the observed relative frequencies are calculated as the average of the observed values, which corresponds to the left-hand side of equation (2.2) conditioning on the partition. The observed relative frequencies are then typically plotted against the average forecast within the bin (or the midpoints) resulting in the so-called *calibration curve*. If the calibration curve is close to the diagonal, equation (2.2) is approximately satisfied and the forecasts are called calibrated or reliable. Figure 2.2 illustrates typical kinds of systematic miscalibration such as an S-shaped curve for underconfidence, that is, the probabilities are too close to the center, or an inverse S-shaped curve for overconfidence, that is, the probabilities are too extreme. Biased probability forecasts result in a calibration curve that is below (above) the diagonal for a positive (negative) bias meaning that the probability forecasts are larger (smaller) than the observed frequencies.

A major drawback of reliability diagrams is that the bins have to be chosen subjectively, typically a number of 10 to 20 equidistant bins are used, and different choices may result in diverse shapes of calibration curves leading to contrary conclusions. In a recent work, Dimitriadis et al. (2021) developed a novel approach that automatically chooses an optimal binning based on the *pool-adjacent-violators* (PAV; de Leeuw et al., 2009) algorithm. In this work, such as in Figure 2.2, we will utilize this approach, which is referred to as CORP (Consistent, Optimal, Reproducible, PAV-based) approach.

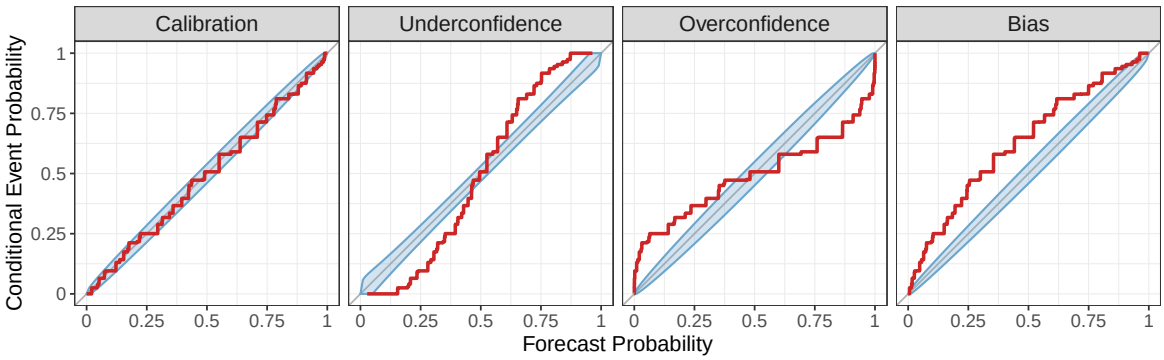

Figure 2.2: Typical shapes of reliability diagrams. From left to right, we observe calibration, underconfidence, overconfidence and a bias. The blue regions show consistency bands at the 90% level under the hypothesis of calibration. The diagrams are based on 10,000 observations drawn from a binomial distribution with probability $p_0 = \Phi(a)$, where a is standard normal and known to all forecasters. The calibrated forecast is given by p_0, the underconfident by $p_1 = \Phi(0.5a)$, the overconfident by $p_2 = \Phi(2a)$ and the biased by $p_3 = \Phi(a - 0.5)$.

2.2 Forecast verification

PIT histograms and reliability diagrams are two tools to qualitatively assess the calibration of a probabilistic forecast. Quantitatively, calibration can be assessed based on the empirical coverage of a PI, sharpness by measures of the spread of the distribution such as the length of a PI. However, we did not provide tools for a simultaneous evaluation of calibration and sharpness. In the following, we will present a principled way to simultaneously assess calibration and sharpness of probabilistic forecasts in order to compare and rank competing forecasters following Gneiting and Raftery (2007).

2.2.1 Proper scoring rules

In the quantitative assessment of the predictive performance, we aim to summarize the goodness of a probabilistic forecast with a numerical score, for a given observation. A *scoring rule* is any extended real-valued function

$$S : \mathcal{F} \times \mathbb{R} \longrightarrow \bar{\mathbb{R}}, \ (F, y) \longmapsto S(F, y) \tag{2.3}$$

such that $S(F, \cdot)$ is $\mathcal{F}$-quasi-integrable for all $F \in \mathcal{F}$, where $\mathcal{F}$ is a suitable class of probability measures. The score $S(F, y)$ is negatively oriented and therefore interpreted as a penalty. A scoring rule S is *proper* relative to the class $\mathcal{F}$ if

$$\mathbb{E}_G[S(G, Y)] \leq \mathbb{E}_G[S(F, Y)] \quad \forall F, G \in \mathcal{F}. \tag{2.4}$$

It is *strictly proper* if equation (2.4) holds with equality only if $F = G$. Propriety is a critical requirement for scoring rules as it prevents the forecaster from hedging its prediction and thereby encourages honest forecasting. Given a set of forecast-observation pairs $\{(F_1, y_1), \ldots, (F_n, y_n)\}$, the mean score $\overline{S}_n^F = \frac{1}{n}\sum_{j=1}^n S(F_j, y_j)$ is calculated in practice to compare and rank different forecasting methods.

To assess the improvement of competing forecasting methods based on a proper scoring rule with respect to a benchmark and allow comparability over different underlying datasets, we can calculate the associated *skill score*. Let $\overline{S}_n^F$ denote the mean score of the forecasting method of interest over a given dataset, $\overline{S}_n^{F_0}$ the corresponding mean score of the benchmark forecast, and $\overline{S}_n^{F^*}$ that of the (typically hypothetical) optimal forecast. The associated skill score SS_F is then calculated via

$$\mathrm{SS}_F = \frac{\overline{S}_n^{F_0} - \overline{S}_n^F}{\overline{S}_n^{F_0} - \overline{S}_n^{F^*}}, \tag{2.5}$$

and simplifies to

$$\mathrm{SS}_F = 1 - \frac{\overline{S}_n^F}{\overline{S}_n^{F_0}} \tag{2.6}$$

if $\overline{S}_n^{F^*} = 0$. In contrast to proper scoring rules, skill scores are positively oriented with 1 indicating optimal predictive performance, 0 no improvement over the benchmark and a negative skill a decrease in performance. Note that skill scores itself are not proper scoring rules, even if the underlying scoring rule is proper.

Comparing two forecasting methods F and G, we can perform a statistical test of equal predictive performance via the *Diebold-Mariano test* (DM test; Diebold and Mariano, 1995). If the forecast cases are independent, the corresponding test statistic is given by

$$t_n = \sqrt{n}\frac{\overline{S}_n^F - \overline{S}_n^G}{\hat{\sigma}_n}, \tag{2.7}$$

where

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{j=1}^n \left(S(F_j, y_j) - S(G_j, y_j)\right)^2, \tag{2.8}$$

and the forecast-observation triples (F_j, G_j, y_j), $j = 1, \ldots, n$, generate a test set of size n. Subject to weak regularity conditions, the test statistic is asymptotically standard normal under the null hypothesis of equal predictive performance, that is, $\overline{S}_n^F - \overline{S}_n^G = 0$. The null hypothesis is rejected for large (absolute) values of t_n, where F is preferred if the test statistic is negative, and G if it is positive.

In Chapter 5, we want to draw conclusions on the statistical significance of score differences on a global level based on a set of DM tests performed on a local level, meaning at different locations and forecast horizons, which confronts us with the problem of assessing significance to a set of multiple tests. Following suggestions of Wilks (2016), we apply a *Benjamini-Hochberg procedure* (Benjamini and Hochberg, 1995) that allows to account for multiple testing and to control the false discovery rate, which we choose to be $\alpha = 0.05$. Given the ordered p-values $\{p_{(1)}, \dots, p_{(M)}\}$ of M hypothesis tests, a threshold p-value is determined via

$$p^* = p_{(i^*)}, \quad \text{where} \quad i^* = \min\left\{ i = 1, \dots, M : p_{(i)} \leq \alpha \cdot \frac{i}{M} \right\}. \tag{2.9}$$

This threshold p^* is then used to decide whether the null hypothesises of the individual tests are rejected.

In the following, we will present some prominent proper scoring rules that are relevant throughout this work. The *continuous ranked probability score* (CRPS; Matheson and Winkler, 1976) is one of the most prominent strictly proper scoring rules in atmospheric sciences and given by

$$\begin{aligned} \mathrm{CRPS}(F, y) &= \int_{-\infty}^{\infty} \left(F(z) - \mathbb{1}\{y \leq z\}\right)^2 dz, & (2.10) \\ &= \mathbb{E}_F \left|Y - y\right| - \frac{1}{2}\mathbb{E}_F \left|Y - Y'\right|, & (2.11) \end{aligned}$$

for observations $y \in \mathbb{R}$ and forecast distributions F with finite first moment, where Y and Y' are independent random variables with CDF F. The CRPS is given in the same unit as the observation and generalizes to the absolute error in case of a deterministic forecast. The integral can be calculated analytically for a wide range of forecast distributions, e.g., for a normal distribution (e.g., Jordan et al., 2019). The *energy score* (ES; Gneiting and Raftery, 2007) is an extension of the CRPS towards multivariate forecasts based on the expectation representation in equation (2.11):

$$\mathrm{ES}(F, \boldsymbol{y}) = \mathbb{E}_F \|\boldsymbol{Y} - \boldsymbol{y}\| - \frac{1}{2}\mathbb{E}_F \|\boldsymbol{Y} - \boldsymbol{Y'}\|, \quad \boldsymbol{y} \in \mathbb{R}^d, \tag{2.12}$$

where d is the dimension of the variable of interest, $\|\cdot\|$ the Euclidean norm on $\mathbb{R}^d$ and F refers to a multivariate forecast CDF on $\mathbb{R}^d$ with $\mathbb{E}_F\|\boldsymbol{Y}\| < \infty$, where $\boldsymbol{Y}$ and $\boldsymbol{Y'}$ are independent random variables with CDF F. In practice, the ES is evaluated empirically based on a sample of the forecast distribution (Jordan et al., 2019). Another popular strictly proper scoring rule is the *logarithmic score* (LogS; Good, 1952) or *ignorance score*

$$\mathrm{LogS}(f, y) = -\log\left(f(y)\right), \quad y \in \mathbb{R}, \tag{2.13}$$

where f denotes the probability density function (PDF) of the probabilistic forecast.

In addition to these proper scoring rules, we introduce two proper scoring rules that are used for probability forecasts. The *Brier score* (BS; Brier, 1950) is defined by

$$\mathrm{BS}(p, y) := (y - p)^2, \quad p \in [0, 1], \quad y \in \{0, 1\}. \tag{2.14}$$

In case of a continuous target variable, the BS can be used to assess forecasts of threshold exceedance derived from the predictive distribution. Given a threshold $z \in \mathbb{R}$, the probability forecast p is given by $F(z)$ and the binary observation y by $\mathbb{1}\{y \leq z\}$.[2] Note that the CRPS is equivalent to the integral over the BS for threshold exceedance of the integration variable. The LogS for probability forecasts, also referred to as *log-loss*, is defined by

$$\mathrm{LogS}(p, y) := -(1 - y) \log(1 - p) - y \log p, \quad p \in [0, 1], \quad y \in \{0, 1\}, \tag{2.15}$$

where $\log(0) := -\infty$.

Strictly proper scoring rules are not only used for the comparison of probabilistic forecasts but also for parameter estimation, which is then referred to as *optimum score estimation* (Gneiting and Raftery, 2007). Let $F(\boldsymbol{x}; \theta)$ be a parametric forecast distribution dependent on the predictor variables $\boldsymbol{x} \in \mathbb{R}^p$ and parameter (vector) $\theta \in \Theta$, where $\Theta \subseteq \mathbb{R}^d$ is the parameter space. We can estimate the optimal parameter (vector) θ by minimizing the mean score of a strictly proper scoring rule S, that is,

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{j=1}^{n} S(F(\boldsymbol{x}_j; \theta), y_j), \tag{2.16}$$

where $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ denotes a training set of size n. Note that minimizing the LogS is equivalent to maximum likelihood estimation (MLE).

2.2.2 Consistent scoring functions

In practical situations, a probabilistic forecast might be reduced to a single value via a statistical functional such as the mean, median or a quantile. For such cases, so-called consistent scoring functions provide useful techniques for forecast evaluation and induce corresponding proper scoring rules (Gneiting, 2011). A *scoring function* $s : \mathbb{R} \times \mathbb{R} \longmapsto [0, \infty)$ is called *consistent* for the functional T relative to the class $\mathcal{F}$ of probability measures if

$$\mathbb{E}_F[s(t, Y)] \leq \mathbb{E}_F[s(x, Y)] \quad \forall F \in \mathcal{F},\, t \in T(F),\, y \in \mathbb{R}. \tag{2.17}$$

In particular, we use the *quantile score* (QS) or *pinball loss* at level $\tau \in (0, 1)$ that is consistent for the quantile q_τ at level τ

$$\rho_\tau(q_\tau, y) = (q_\tau - y)(\mathbb{1}\{y \leq q_\tau\} - \tau), \quad y \in \mathbb{R}. \tag{2.18}$$

[2] Note that this is equivalent to using $p = 1 - F(z)$ and $y = \mathbb{1}\{y > z\}$ but a more convenient notation.

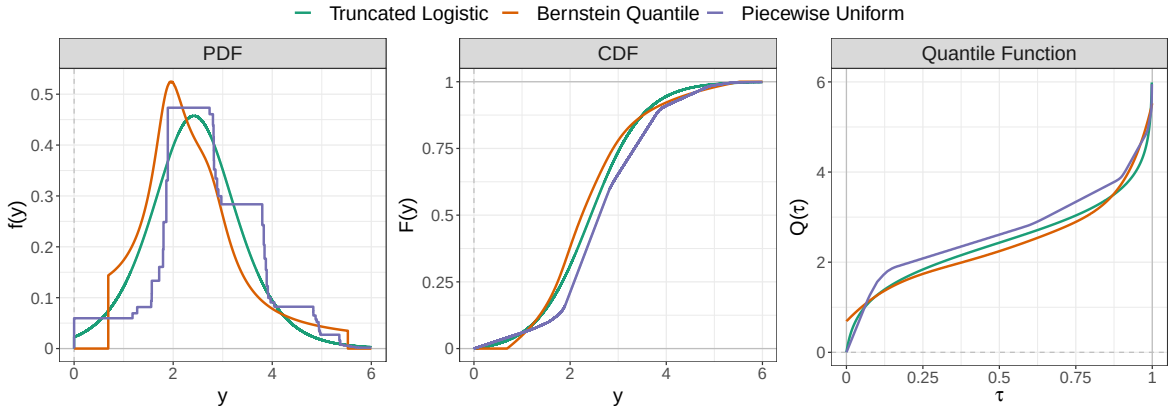

Figure 2.3: PDF, CDF and quantile function (left to right) of exemplary forecasts from a zero-truncated logistic distribution, a Bernstein quantile function of degree 12 and a piecewise uniform distribution based on 55 bins. The forecasts are taken from the case study on wind gust forecasting in Section 5.3. Note that both the target variable and the forecasts are positive.

In case of the median, the QS reduces to the *absolute error* (AE)

$$\mathrm{AE}(q_{0.5}, y) = |q_{0.5} - y|, \quad y \in \mathbb{R}, \tag{2.19}$$

which yields the mean absolute error (MAE). The difference between a point forecast and the observation is not a (consistent) scoring function, but will be used to analyze the bias of the median forecast. The QS also yields the CRPS via

$$\mathrm{CRPS}(F, y) = \int_0^1 2\rho_\tau(q_\tau, y)d\tau, \quad y \in \mathbb{R}, \tag{2.20}$$

where q_τ denotes the quantile at level τ (Gneiting and Ranjan, 2011). Further, we use the *squared error* (SE) that is consistent for the mean $\overline{x}$

$$\mathrm{SE}(\overline{x}, y) = (\overline{x} - y)^2, \quad y \in \mathbb{R}. \tag{2.21}$$

In practice, the mean squared error is typically transformed via the square root to the *root mean squared error* (RMSE) such that the score has the same unit as the observation.

2.3 Exemplary types of forecast distributions

As indicated in the previous sections, probabilistic forecasts may take different forms such as a fully specified predictive distribution, an ensemble forecast, a PI or a set of quantiles. Here, we introduce different types of predictive distributions that are used throughout this work.

2.3.1 TRUNCATED AND CENSORED LOGISTIC DISTRIBUTION

Distributions such as the normal and logistic distribution are popular for modeling variables in statistics due to their appealing properties. However, if a variable is not real-valued but instead positive or has a mixed discrete-continuous distribution, such as wind gusts or solar irradiance, which we focus on in Section 5, these distributions are not suitable. Still, one can utilize them by modification towards the variable of interest, for which truncation and censoring are two options.

We will present both of these options based on the logistic distribution, which is given by the CDF

$$F\left(y;\mu,\sigma\right) = \left(1+\exp\left(-z_y\right)\right)^{-1} \quad \text{with} \quad z_y := \frac{y-\mu}{\sigma}, \quad y \in \mathbb{R}, \tag{2.22}$$

where $\mu \in \mathbb{R}$ denotes the location and $\sigma > 0$ the scale parameter. Note that, in general, truncation and censoring can be applied to other real-valued distributions such as the normal distribution as well.

First, we present the concept of (left-)truncation, which is used to cut the support to a subset thereof, such as the positive halfaxis. In general, let F be a CDF and $t \in \mathbb{R}$ the truncation threshold. Then, the distribution (left-)truncated in t or t-truncated distribution with support (t,∞) is defined by the CDF

$$F_t(y) := \frac{F(y)-F(t)}{1-F(t)}, \quad y > t. \tag{2.23}$$

From this formula, we can derive the associated PDF

$$f_t(y) = \frac{f(y)}{1-F(t)}, \quad y > t, \tag{2.24}$$

and, under the assumption of strict monotonicity, the quantile function

$$F_t^{-1}(\tau) = F^{-1}\left(F(t)+\tau\left(1-F(t)\right)\right), \quad \tau \in (0,1). \tag{2.25}$$

In a nutshell, truncation cuts off the distribution at t and allocates the probability mass that was cut off proportionally to the remainder of the distribution. For the logistic distribution left-truncated in zero, we obtain the following formula for the CDF

$$F_0\left(y;\mu,\sigma\right) := \frac{F\left(y;\mu,\sigma\right)-F\left(0;\mu,\sigma\right)}{1-F\left(0;\mu,\sigma\right)} = \frac{\exp\left(z_y\right)-\exp\left(z_0\right)}{1+\exp\left(z_y\right)}, \quad y > 0, \tag{2.26}$$

where $\mu \in \mathbb{R}$ and $\sigma > 0$ are again referred to as location and scale parameter, respectively, and z_y is defined as in equation (2.22). Note that in contrast to the logistic distribution, the truncated variant is not symmetric and the location parameter is not identical to the mean and median of the distribution. In particular, negative location parameters are still valid.

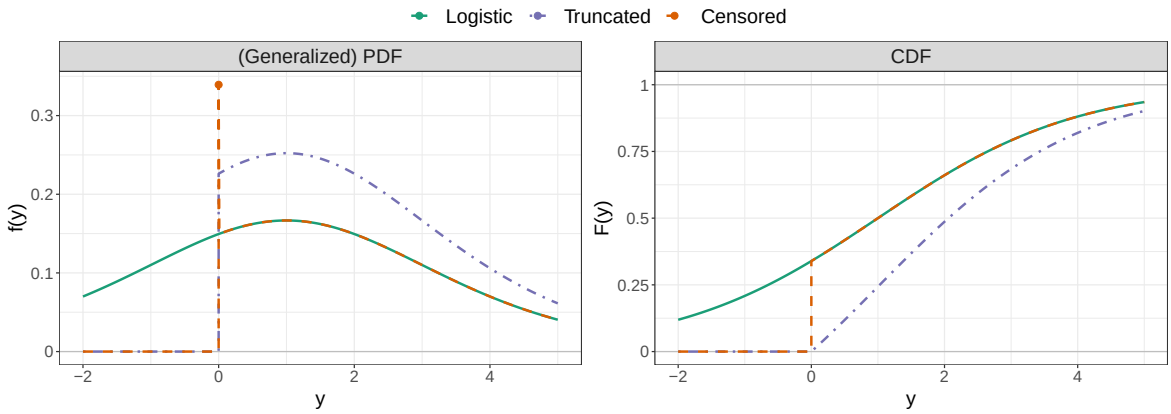

Figure 2.4: (Generalized) PDF (left) and CDF (right) of a logistic distribution with location parameter $\mu = 1$ and scale parameter $\sigma = 1.5$ as well as the zero-truncated and zero-censored variant thereof (adapted from Figure 1 in Schulz et al., 2021).

The mean of the zero-truncated logistic distribution is given by

$$\frac{\mu - \sigma \log\left(1 - F\left(z_0\right)\right)}{1 - F\left(z_0\right)} \tag{2.27}$$

and the median by

$$\mu + \sigma \log\left(1 + 2\exp\left(z_0\right)\right). \tag{2.28}$$

Recalling the PDF of the truncated distribution in equation (2.24), we note that the mode of the truncated logistic distribution is given by μ for $\mu > 0$ and 0 otherwise. A forecast in the form of a zero-truncated logistic distribution is simply given by the location and scale parameter that define the full predictive distribution. Further, a mean, median and quantile forecast can be derived using equations (2.27), (2.28) and (2.25). While the LogS is straightforward to calculate by plugging in the PDF of the zero-truncated logistic distribution, the CRPS can be calculated via an analytic formula that has been derived by Scheuerer and Möller (2015). Figure 2.3 shows an exemplary forecast in the form of a zero-truncated logistic distribution that is included in the case study on wind gust forecasting in Section 5.3, whereas Figure 2.4 illustrates the effect of truncation for a logistic distribution.

While truncation distributes the probability mass that was cut off along the remaining support, censoring assigns it as a point mass to the threshold resulting in a mixed discrete-continuous distribution. Let F be a CDF and $c \in \mathbb{R}$ the censoring threshold, then the CDF of a distribution (left-)censored in c or c-censored distribution is given by

$$F_c(y) := \begin{cases} 0, & y < c, \\ F(y), & y \geq c. \end{cases} \tag{2.29}$$

The (generalized)[3] PDF is given by

$$f_c(y) = F(c) \cdot \mathbb{1}\{y = c\} + f(c) \cdot \mathbb{1}\{y > c\}, \quad y \in \mathbb{R}, \tag{2.30}$$

and, under the assumption of strict monotonicity, the quantile function by

$$F_c^{-1}(\tau) = \begin{cases} c, & \tau < F(c), \\ F^{-1}(\tau), & \tau \geq F(c). \end{cases} \tag{2.31}$$

Zero-censoring a logistic distribution has similar consequences as zero-truncation in that the censored variant is not symmetric and the location parameter is not identical to the mean anymore, which is given by

$$\mu + \sigma \log\left(1 + \exp\left(-z_0\right)\right). \tag{2.32}$$

Further, negative location parameter are valid choices. In contrast to truncation, the median is still given by the location parameter for nonnegative choices and elsewise by 0, as the quantile function in equation (2.31) shows. The same applies for the mode. The full predictive distribution of a zero-censored distribution is given by the location and scale parameter, point forecasts can be readily derived. The LogS can be calculated based on the generalized PDF in equation (2.30), an analytic formula for the CRPS was derived by Taillardat et al. (2016). A more general formula for the CRPS including both truncation and censoring of a logistic distribution is given by Jordan et al. (2019). The effects of truncation and censoring of a logistic distribution are illustrated and compared in Figure 2.4.

2.3.2 Bernstein quantile function

The second exemplary type of forecast distribution is based on the class of *Bernstein polynomials* that have the property to approximate any continuous function on the unit interval (Bustamante, 2017). Due to this and other appealing properties, Bernstein polynomials are often used to model quantile functions (e.g., Wang and Ghosh, 2012). Hence, the idea of this type of forecast distribution is to model the quantile function via a Bernstein polynomial, which is a linear combination of Bernstein basis polynomials (Bremnes, 2020). We refer to the forecast via the *Bernstein quantile function*, which is given by

$$Q\left(\tau\right) := \sum_{l=0}^{d} \alpha_l B_{l,d}(\tau), \quad \tau \in [0, 1], \tag{2.33}$$

[3] Generalized PDFs are a generalization from continuous to mixed discrete-continuous distributions.

with basis coefficients $\alpha_0 \leq \cdots \leq \alpha_d$, where

$$B_{l,d}(\tau) = \binom{d}{l} \tau^l (1-\tau)^{d-l}, \quad l = 0, \ldots, d, \tag{2.34}$$

are the Bernstein basis polynomials of degree $d \in \mathbb{N}$.

A critical requirement is that the coefficients are nondecreasing, because this implies that the quantile function is also nondecreasing. If the coefficients are strictly increasing, the same holds for the quantile function. Following Wang and Ghosh (2012), the derivative of equation (2.33) is given by

$$Q'(\tau) = d \sum_{l=0}^{d-1} (\alpha_{l+1} - \alpha_l) B_{l,d-1}(\tau). \tag{2.35}$$

The Bernstein polynomials are positive on the open unit interval, that is, $B_{l,d}(\tau) > 0$ for $\tau \in (0,1)$ and $l = 0, \ldots, d$, therefore the derivative is nonnegative (positive) for $\tau \in (0,1)$ if the coefficients are (strictly) increasing.

The mean of the Bernstein forecast distribution can be calculated using the following property of the Bernstein basis polynomials (Bustamante, 2017):

$$\int_0^1 B_{ld}(\tau)\, d\tau = \frac{1}{d+1}, \quad l = 0, \ldots, d. \tag{2.36}$$

Based on this property, the mean of a random variable with a Bernstein quantile function is calculated via

$$\int_0^1 Q(\tau)\, d\tau = \sum_{l=0}^{d} \alpha_l \int_0^1 B_{l,d}(\tau) d\tau = \frac{1}{d+1} \sum_{l=0}^{d} \alpha_l. \tag{2.37}$$

Further, the support of the distribution is given by $[\alpha_0, \alpha_d]$, which directly follows from $Q(0) = \alpha_0$ and $Q(1) = \alpha_d$. Hence, a positive first coefficient results in a positive support of the forecast distribution. Figure 2.3 illustrates an exemplary forecast from a Bernstein quantile function of degree $d = 12$ taken from the case study on wind gust prediction in Section 5.3. We can observe that the forecast distribution is smooth and more flexible than that of a truncated logistic distribution, where the shape of the distribution is predetermined. Further, the finite support becomes apparent for the PDF of the forecast distribution.

From the point of forecast verification, one downside of the Bernstein quantile function is that in general the CDF and PDF cannot be calculated analytically. Therefore, the LogS and the CRPS cannot be calculated analytically. Hence, we approximate both scores based on a set of equidistant quantiles, which is also used to calculate the PIT values for the forecast evaluation.

2.3.3 PIECEWISE UNIFORM DISTRIBUTION

The motivation for the last exemplary type of forecast distribution is to approximate the predictive density with a histogram, which corresponds to the idea of transforming the probabilistic forecasting problem into a classification task. Here, the histogram is based on a partition of the observation range, where each bin can be interpreted as a class, for which the bin frequency is the associated class probability. On the other hand, a histogram can be interpreted as a piecewise constant function that yields a density (up to a scaling factor). A piecewise constant PDF corresponds to a piecewise uniform distribution, where we implicitly assume a uniform distribution within each bin. Hence, we can create a predictive distribution by solving a discrete classification problem.

To formally introduce the piecewise uniform distribution, let N be the number of histogram bins and $b_0 < \cdots < b_N$ the edges of the bins $I_l = [b_{l-1}, b_l)$ with probabilities p_l, $l = 1, \ldots, N$, where it holds that $\sum_{l=1}^{N} p_l = 1$. For completeness, we define $b_{-1} = -\infty$ and $b_{N+1} = \infty$, which yield the bins $I_0 = (-\infty, b_0)$ and $I_{N+1} = [b_N, \infty)$, to which we both assign probability $p_0 = p_{N+1} = 0$, meaning that we only consider partitions that yield a distribution with finite support. Further, the projection of $y \in \mathbb{R}$ to the interval $[a, b)$ is given by

$$(y|_{[a,b)}) := \max\{a, \min\{y, b\}\} = \begin{cases} a, & y < a, \\ y, & a \leq y < b, \\ b, & b \leq y. \end{cases} \tag{2.38}$$

Now, we can define the bin in which a realization $y \in \mathbb{R}$ falls via

$$\begin{aligned} \kappa := \kappa(y) &:= \sum_{l=0}^{N+1} l \cdot \mathbb{1}\{y \in I_l\} \\ &= \min\{l : y < b_l,\, 0 \leq l \leq N+1\} = \max\{l : b_{l-1} \leq y,\, 0 \leq l \leq N+1\}. \end{aligned} \tag{2.39}$$

As mentioned before, the PDF of the corresponding probabilistic forecast is a piecewise constant function given by

$$f(y) = \sum_{l=1}^{N} \frac{p_l}{b_l - b_{l-1}} \cdot \mathbb{1}\{y \in I_l\} = \frac{p_\kappa}{b_\kappa - b_{\kappa-1}}, \quad y \in \mathbb{R}. \tag{2.40}$$

Note that the PDF reduces to the values of the associated bin probability when the bin lengths are equidistant with length 1, that is, $b_l - b_{l-1} = 1$ for $l = 1, \ldots, N$.

Due to the fact that the PDF is defined as a histogram, the CDF is a piecewise linear function given by

$$F(y) = \sum_{l=1}^{N} p_l \left(\frac{y - b_{l-1}}{b_l - b_{l-1}} \cdot \mathbb{1}\{y \in I_l\} + \mathbb{1}\{b_l \leq y\} \right) = \sum_{l=1}^{N} p_l \frac{(y|_{I_l}) - b_{l-1}}{b_l - b_{l-1}} \cdot \mathbb{1}\{b_{l-1} \leq y\}, \tag{2.41}$$

for $y \in \mathbb{R}$ or, equivalently, by

$$F(y) = p^*_{(\kappa-1)} + p_\kappa \frac{y - b_{\kappa-1}}{b_\kappa - b_{\kappa-1}}, \quad y \in \mathbb{R}, \tag{2.42}$$

where $p^*_{(l)} := \sum_{j=0}^{l} p_j$ is the accumulated probability up to the lth bin for $l = 0, \ldots, N$.

Hence, the quantile function is also a piecewise linear function, for which we characterize the bins in terms of the accumulated probabilities via $P^*_l := [p^*_{(l-1)}, p^*_{(l)})$ for $l = 1, \ldots, N$ with $P^*_l = \emptyset$ for $p_l = 0$. Using the characterization of the bins in terms of the accumulated probability, we define the bin of a certain quantile level $\tau \in (0, 1)$ by

$$\begin{aligned} \kappa := \kappa(\tau) &:= \sum_{l=1}^{N} l \cdot \mathbb{1}\{\tau \in P^*_l\} \\ &= \min\{l : \tau < p^*_{(l)},\, 1 \le l \le N\} = \max\{l : p^*_{(l-1)} \le \tau,\, 1 \le l \le N\} \end{aligned} \tag{2.43}$$

with $\kappa \in \{l = 1, \ldots, N : p_l > 0\}$. Based on the bin κ, the quantile function is given by

$$Q(\tau) = b_{\kappa-1} + (b_\kappa - b_{\kappa-1}) \frac{\tau - p^*_{(\kappa-1)}}{p_\kappa}, \quad \tau \in (0, 1)\,. \tag{2.44}$$

Note that $p_\kappa > 0$ and that the denominator results from the fact that $p_l = p^*_{(l)} - p^*_{(l-1)}$ for $l = 1, \ldots, N$. Analogous to equation (2.41), we can formulate the quantile function alternatively via

$$Q(\tau) = b_0 + \sum_{l=1}^{N} (b_l - b_{l-1}) \left(\frac{\tau - p^*_{(l-1)}}{p_l} \cdot \mathbb{1}\{\tau \in P^*_l\} + \mathbb{1}\{p^*_{(l)} \le \tau\} \right) \tag{2.45}$$

$$= b_0 + \sum_{l=1}^{N} (b_l - b_{l-1}) \frac{(\tau|_{P^*_l}) - p^*_{(l-1)}}{p_l} \cdot \mathbb{1}\{p^*_{(l-1)} \le \tau\}, \tag{2.46}$$

where $\tau \in (0, 1)$. In equation (2.45), we can neglect the special case of division by $p_l = 0$ for $l = 1, \ldots, N$, as we have $P^*_l = \emptyset$ and $\mathbb{1}\{\tau \in P^*_l\} = 0$ in that case. For equation (2.45), we define $0/0 = 1$. Note that the predictive distribution is defined by the binning scheme and the corresponding bin probabilities. Figure 2.3 shows an exemplary forecast of a piecewise uniform distribution with $N = 55$ nonequidistant bins from the case study on probabilistic wind gust forecasting in Section 5.3, where the piecewise constant structure of the PDF is apparent.

The structure of the predictive distribution allows for a straightforward computation of the quantities of interest such as the CRPS, LogS or mean. The mean of a piecewise uniform distribution can be calculated as

$$\int_{-\infty}^{\infty} x f(x) dx = \frac{1}{2} \sum_{l=1}^{N} p_l \,(b_{l-1} + b_l)\,. \tag{2.47}$$

The CRPS of the piecewise uniform distribution can be calculated using the CRPS of an uniform distribution with point masses on the left and right boundaries, for which an analytic solution is given by Jordan et al. (2019).

Proposition 2.1. *Let F be the predictive CDF as defined in equation (2.41) and $y \in \mathbb{R}$. With the definition of $I := \bigcup_{l=1}^{N} I_l = [b_0, b_N)$, and the uniform distribution on I_l with point masses $p^*_{(l-1)}$ on the lower and $1 - p^*_{(l)}$ on the upper boundary given by the CDF*

$$F_l(y) := \begin{cases} 0, & y < b_{l-1}, \\ p^*_{(l-1)} + p_l \dfrac{y - b_{l-1}}{b_l - b_{l-1}}, & b_{l-1} \leq y < b_l, \\ 1, & b_l \leq y, \end{cases} \tag{2.48}$$

the CRPS can be calculated via

$$\mathrm{CRPS}(F, y) = |y - (y|_I)| + \sum_{l=1}^{N} \mathrm{CRPS}\left(F_l, (y|_{I_l})\right). \tag{2.49}$$

Note that $|y - (y|_I)| = 0$ for $y \in I$ and $(y|_{I_l}) = y$ for $y \in I_l$. Figure 2.5 illustrates the CRPS of a piecewise uniform distribution and explains why we need to transform the observations via equation (2.38) for the individual CRPS calculations corresponding to the bins.

Further, note that the CRPS formula in equation (2.49) is a special case of a result from Jordan (2016, Proposition 6.1(d)), which states that the CRPS of a distribution can be split into the CRPSs of censored distributions on a pairwise disjoint cover of the real line. Using the cover $\mathcal{I} = (I_0, \ldots, I_{N+1})$ and the CDF in equations (2.41) and (2.42), we obtain the CRPS in equation (2.49). Here, we provide a proof for the CRPS formula, independent of the result in Jordan (2016).[4]

Proof. Let $y \in \mathbb{R}$ and F be the CDF of a piecewise uniform distribution. Further, we define $h(z; y, G) := (G(z) - \mathbb{1}\{y \leq z\})^2$ for $z \in \mathbb{R}$ and a CDF G. First, we state the following properties for $l = 1, \ldots, N$:

(A1) $F_l(z) = F(z)$ for $z \in I_l$.

(A2) $\mathbb{1}\{(y|_{I_l}) \leq z\} = \mathbb{1}\{y \leq z\}$ for $z \in I_l$, as $(y|_{I_l}) = y$ for $y \in I_l$, $(y|_{I_l}) = b_{l-1}$ for $y < b_{l-1} \leq z$ and $(y|_{I_l}) = b_l$ for $y \geq b_l > z$.

(A3) $F_l(z) = 0 = \mathbb{1}\{(y|_{I_l}) \leq z\}$ for $z < b_{l-1} \leq (y|_{I_l})$.

(A4) $F_l(z) = 1 = \mathbb{1}\{(y|_{I_l}) \leq z\}$ for $z \geq b_l \geq (y|_{I_l})$.

(A5) From (A3) and (A4), it follows that $h(z; (y|_{I_l}), F_l) = 0$ for $z \notin I_l$.

[4] Jordan (2016) concluded that the result "follow[s] straightforwardly from the threshold decomposition" (equation (2.10)).

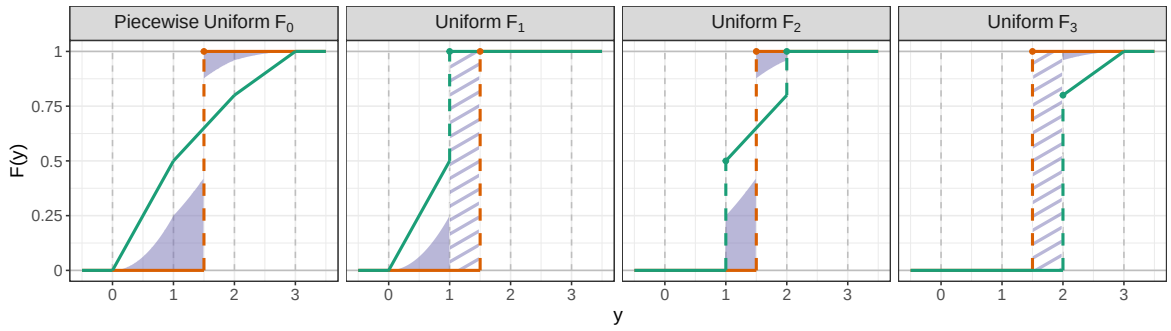

Figure 2.5: Illustration of the CRPS of a piecewise uniform distribution F_0 based on the bin edges $(b_0, b_1, b_2, b_3) = (0, 1, 2, 3)$ and bin probabilities $(p_1, p_2, p_3) = (0.5, 0.3, 0.2)$, together with the observation $y = 1.5$. The left panel shows the CDF F_0 (green) and the indicator function based on y (orange), together with the purple area that is equivalent to the CRPS. Instead of F_0, the following panels show the CDFs F_1, F_2 and F_3 of uniform distributions with point masses on the edges, as defined in equation (2.48). The striped, purple area corresponds to the additional value when using y instead of $(y|_{I_l})$ for the calculation of the individual CRPS values in equation (2.49).

Next, we show that

$$\int_{-\infty}^{b_0} h(z; y, F)dz + \int_{b_N}^{\infty} h(z; y, F)dz = |y - (y|_I)| \,. \tag{2.55}$$

We have

$$\begin{aligned} \int_{-\infty}^{b_0} h(z; y, F)dz + \int_{b_N}^{\infty} h(z; y, F)dz &= \int_{-\infty}^{b_0} (0 - \mathbb{1}\{y \leq z\})^2 dz + \int_{b_N}^{\infty} (1 - \mathbb{1}\{y \leq z\})^2 dz \\ &= \int_{-\infty}^{b_0} \mathbb{1}\{y \leq z\}dz + \int_{b_N}^{\infty} \mathbb{1}\{y > z\}dz \end{aligned}$$

We consider the three following cases. For $y < b_0$, we have $(y|_I) = b_0$ and obtain

$$\int_{-\infty}^{b_0} \mathbb{1}\{y \leq z\}dz + \int_{b_N}^{\infty} \underbrace{\mathbb{1}\{y > z\}}_{=0} dz = \int_{-\infty}^{y} 0dz + \int_{y}^{b_0} 1dz = b_0 - y = y_I - y = |y - (y|_I)| \,.$$

For $y \in I$, we have $(y|_I) = y$ and obtain

$$\int_{-\infty}^{b_0} \underbrace{\mathbb{1}\{y \leq z\}}_{=0} dz + \int_{b_N}^{\infty} \underbrace{\mathbb{1}\{y > z\}}_{=0} dz = 0 = y - y_I = |y - (y|_I)| \,.$$

For $y \geq b_N$, we have $(y|_I) = b_N$ and obtain

$$\int_{-\infty}^{b_0} \underbrace{\mathbb{1}\{y \leq z\}}_{=0} dz + \int_{b_N}^{\infty} \mathbb{1}\{y > z\}dz = \int_{b_N}^{y} 1dz + \int_{y}^{\infty} 0dz = y - b_N = y - (y|_I) = |y - (y|_I)| \,.$$

At last, we calculate the CRPS of the piecewise uniform distribution:

$$\begin{aligned}
\mathrm{CRPS}(F, y) &= \int_{-\infty}^{\infty} h\left(z; y, F\right) dz \\
&= \int_{-\infty}^{b_0} h\left(z; y, F\right) dz + \sum_{l=1}^{N} \int_{I_l} h\left(z; y, F\right) dz + \int_{b_N}^{\infty} h\left(z; y, f\right) dz \\
&\overset{(2.55)}{=} |y - (y|_I)| + \sum_{l=1}^{N} \int_{I_l} h\left(z; y, F\right) dz \\
&\overset{(A1),(A2)}{=} |y - (y|_I)| + \sum_{l=1}^{N} \int_{I_l} h\left(z; (y|_{I_l}), F_l\right) dz \\
&\overset{(A5)}{=} |y - (y|_I)| + \sum_{l=1}^{N} \int_{-\infty}^{\infty} h\left(z; (y|_{I_l}), F_l\right) dz \\
&= |y - (y|_I)| + \sum_{l=1}^{N} \mathrm{CRPS}\left(F_l, (y|_{I_l})\right).
\end{aligned}$$

□

As the LogS is a local scoring rule that only depends on the value the predictive PDF attains at the observation, the LogS reduces to the simple form

$$\mathrm{LogS}(f, y) = -\log \frac{p_\kappa}{b_\kappa - b_{\kappa-1}} = \log\left(b_\kappa - b_{\kappa-1}\right) - \log\left(p_\kappa\right). \tag{2.56}$$

In the aforementioned case of equidistant bins of length 1, the LogS becomes even simpler:

$$\mathrm{LogS}(f, y) = \log(1) - \log\left(p_\kappa\right) = -\log\left(p_\kappa\right). \tag{2.57}$$

At the end of the section, we again draw a connection to classification problems. The LogS in equation (2.57) is equivalent to the categorical cross-entropy, a loss function frequently used in machine learning (ML) to fit statistical models for classification. Minimizing the LogS with fixed, equidistant bins of length 1 to find class probabilities is therefore equivalent to using the categorical cross-entropy in the discrete classification problem. As the term $\log(b_\kappa - b_{\kappa-1})$ is constant for fixed bins and κ depends on the observation only, MLE for a piecewise uniform distribution is for fixed bins in general equivalent to using the categorical cross-entropy in the associated classification problem.

CHAPTER 3

Aggregating Distribution Forecasts from Deep Ensembles

Motivated by their superior performance on a wide variety of ML tasks, much recent research interest has focused on the use of deep neural networks (NNs) for probabilistic forecasting. Different approaches for obtaining a forecast distribution as the output of an NN have been proposed over the past years, including parametric methods where the NN outputs parameters of a parametric probability distribution such as the truncated logistic distribution described in Section 2.3.1 (Lakshminarayanan et al., 2017; D'Isanto and Polsterer, 2018; Rasp and Lerch, 2018), semiparametric approximations of the quantile function of the forecast distribution such as the Bernstein quantile function described in Section 2.3.2 (Bremnes, 2020) and nonparametric methods where the forecast density is modeled as a histogram such as for the piecewise uniform distribution described in Section 2.3.3 (Gasthaus et al., 2019; Li et al., 2021). To account for the randomness of the training process based on stochastic gradient descent methods, NNs are often run several times from different random initializations. Lakshminarayanan et al. (2017) refer to this simple to implement and readily parallelizable approach as *deep ensembles*. Deep ensembles of NN models for probabilistic forecasting thus yield an ensemble of predictive probability distributions. To provide a final probabilistic forecast, the ensemble of predictive distributions needs to be aggregated to obtain a single forecast distribution.

The problem of combining predictive distributions has been studied extensively in the statistical literature, see Gneiting and Ranjan (2013) and Petropoulos et al. (2022, Section 2.6) for overviews. Combining probabilistic forecasts from different sources has been successfully used in a wide variety applications including economics (Aastveit et al., 2018), epidemiology (Cramer et al., 2022; Taylor and Taylor, 2023), finance (Berkowitz, 2001), signal processing (Koliander et al., 2022) and weather forecasting (Baran and Lerch, 2016, 2018), and constitutes one of the typical components of winning submissions to forecasting competitions (Bojer

and Meldgaard, 2021; Januschowski et al., 2022). On the other hand, forecast combination also forms the theoretical framework of some of the most prominent techniques in ML such as boosting (Freund and Schapire, 1996), bagging (Breiman, 1996) or random forests (RFs; Breiman, 2001), which are based on the idea of building ensembles of learners and combining the associated predictions. Generally, the individual component models (or ensemble members) can be based on entirely distinct modeling approaches, or on a common modeling framework where the model training is subject to different input datasets of other sources of stochasticity. The latter is the case for deep ensembles where the main sources of uncertainty in the estimation are the random initialization of the network parameters and the stochastic gradient descent algorithm for the optimization. For general reviews on ensemble methods in ML, we refer to Dietterich (2000), Zhou et al. (2002) and Ren et al. (2016).

While the arithmetic mean is a powerful and widely accepted method for aggregating single-valued point forecasts, the question how probabilistic forecasts should be combined is more involved and has been a focus of research interest in the literature on statistical forecasting (Gneiting and Ranjan, 2013; Lichtendahl et al., 2013; Petropoulos et al., 2022). We will focus on readily applicable aggregation methods for the combination of probabilistic forecasts from deep ensembles. A widely used approach is the linear aggregation of the forecast distributions, an approach that is often referred to as *linear (opinion) pool* (LP). In the context of deep ensembles, Lakshminarayanan et al. (2017) apply the LP and linearly combine density forecasts. An alternative is given by aggregating the forecast distributions on the scale of quantiles by linearly combining the corresponding quantile functions, an approach that is commonly referred to as *Vincentization* (VI).

The main aim of this chapter is to consolidate findings from the statistical and ML literature on forecast combination and ensembling for probabilistic forecasting. Using theoretical arguments (Sections 3.1 and 3.2), simulation experiments (Section 3.3) and a case study on probabilistic wind gust forecasting (Section 3.4), we systematically investigate and compare aggregation methods for probabilistic forecasts based on deep ensembles, with different ways to characterize the corresponding forecast distributions. In the following, we apply a two-step procedure by first generating an ensemble of probabilistic forecasts and then aggregating them into a single final forecast, which matches the typical workflow of forecast combination from a forecasting perspective. Alternatively, it is also possible to incorporate the aggregation procedure directly into the model estimation (Kim et al., 2021).

The remainder of the chapter starts with Section 3.1 that introduces the forecast aggregation methods. For the three exemplary distributions presented in Section 2.3, which will form the basis for the NNs in Chapter 4, we discuss in Section 3.2 how the different aggregation methods can be used to combine the corresponding predictive distributions of an ensemble of such forecasts. In Section 3.3, we conduct a comprehensive simulation study that is followed up by a case study on probabilistic weather prediction in Section 3.4. Section 3.5 concludes with a discussion.

3.1 Combining predictive distributions

Here, we formally introduce the LP and VI methods for aggregating probabilistic forecasts. Given $K \in \mathbb{N}$ individual probabilistic forecasts we aim to aggregate, we will denote their CDFs by $F_1, \ldots, F_K$ and their quantile functions by $Q_1, \ldots, Q_K$. The aggregation methods introduced below will typically assign weights $w_1, \ldots, w_K$ to the individual forecast distributions.

3.1.1 Linear pool

The most widely used approach for forecast combination is the LP, which is the arithmetic mean of the individual forecasts (Stone, 1961). For probabilistic forecasts, the LP is calculated as the (in our case equally) weighted average of the predictive CDFs and results in a mixture distribution. Equivalently, the LP can be calculated by averaging the PDFs. We define the predictive CDF of the LP via

$$F_w(y) := \sum_{i=1}^{K} w_i F_i(y), \quad y \in \mathbb{R}, \tag{3.1}$$

where $w_i \geq 0$ for $i = 1, \ldots, K$ with $\sum_{i=1}^{K} w_i = 1$. Note that the weights need to sum up to 1 to ensure that F_w yields a valid CDF.

The LP has some appealing theoretical properties and has been the prevalent forecast aggregation method over the last decades.[1] For example, Lakshminarayanan et al. (2017) use the LP to combine density forecasts of multiple NNs introducing the term deep ensembles. However, there are disadvantages to the use of the LP that is known to have suboptimal properties when aggregating probabilities, since a linear combination of probability forecasts results in less sharp and more underconfident forecasts (Ranjan and Gneiting, 2010). Gneiting and Ranjan (2013) extend this result to the general case of predictive distributions by showing that in case of distribution forecasts sharpness decreases and dispersion increases. In particular, a (nontrivial) combination of calibrated forecasts is not calibrated anymore. In the context of deep ensembles, these downsides have also been observed in recent studies (Rahaman and Thiery, 2021; Wu and Gales, 2021).

In our simulation and case study conducted in the following, we apply the aggregation methods to forecasts produced by the same data-generating mechanism based on an ensemble of NNs, which differ only in the random initialization. Therefore, we do not expect systematic differences between the individual forecasts and only consider equally weighted averages. In the following, we will refer to the LP as the equally weighted average given by $w_i = 1/K$ for $i = 1, \ldots, K$ in equation (3.1). Figure 3.1 illustrates the effect of forecast combination via the LP.

[1] For example, Lichtendahl et al. (2013) and Abe et al. (2022) show that the score of the LP forecast is at least as good as the average score of the individual components in terms of different proper scoring rules.

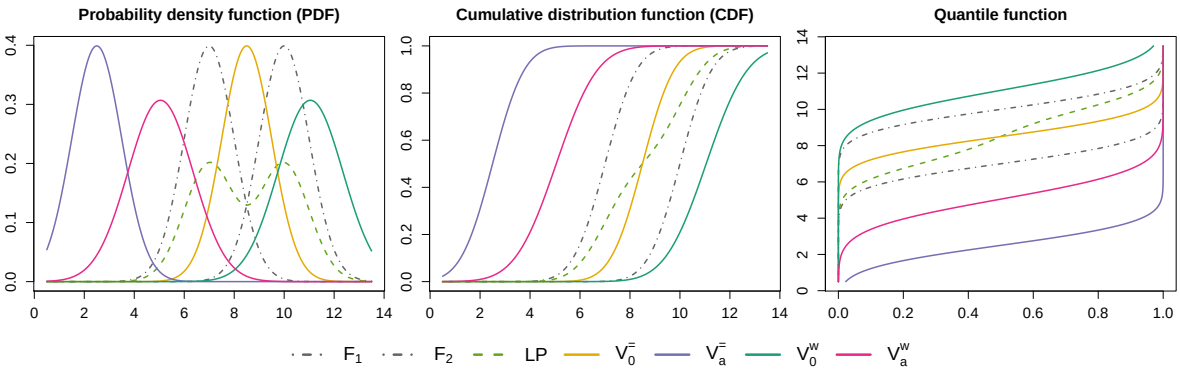

Figure 3.1: PDF, CDF and quantile function of two normally distributed forecasts F_1 and F_2 ($\mu_1 = 7$, $\mu_2 = 10$, $\sigma_1 = \sigma_2 = 1$) together with forecasts aggregated via the methods presented in Section 3.1. $\text{V}_a^=$ and V_a^w use the intercept $a = -6$, V_0^w and V_a^w the weight $w_0 = 0.6$.

3.1.2 VINCENTIZATION

While the LP aggregates the forecasts on a probability scale, VI performs a quantile-based linear aggregation (Vincent, 1912; Ratcliff, 1979; Genest, 1992). We extend the standard VI framework by defining the VI quantile function via

$$Q_w^a(\tau) := a + \sum_{i=1}^{K} w_i Q_i(\tau), \quad \tau \in [0, 1], \tag{3.2}$$

where $a \in \mathbb{R}$ and $w_i \geq 0$ for $i = 1, \ldots, K$.[2] In contrast to the LP, the weights do not need to sum to 1 and only their nonnegativity is required to ensure the monotonicity of the resulting quantile function Q_w^a.[3] Further, a real-valued intercept a is added to the aggregated quantile functions to correct for systematic biases.

As for the LP, we only consider equally weighted averages for VI, that is, $w_i = w_0 > 0$ for $i = 1, \ldots, K$. Given equal weights, we consider four different variants of VI. First, with weights that sum up to 1 and no intercept, that is, $a = 0$ and $w_0 = 1/K$, which is referred to by $\text{V}_0^=$. Similar to the LP, $\text{V}_0^=$ does not require the estimation of any parameters. Further, we consider VI variants where we estimate the parameters a and w_0 both independently (while the other is fixed at the values of $\text{V}_0^=$) and also simultaneously, resulting in the three variants $\text{V}_a^=$ (where $w_0 = 1/K$ and a is estimated), V_0^w (where $a = 0$ and w_0 is estimated) and V_a^w (where both a and w_0 are estimated). The parameters are estimated minimizing the CRPS following the optimum scoring principle. The standard procedure for training ML models where the available data is split into training, validation and test datasets offers a natural

[2] To the best of our knowledge, VI is usually only applied with standardized weights $w_i, i = 1, \ldots, K$, with $\sum_{i=1}^{K} w_i = 1$, and without the intercept a. Exceptions include Wolffram (2021) and related, unpublished simulation experiments by Anja Mühlemann (University of Bern, 2020, personal communication).

[3] Note that in general Q_w^a is not the quantile function corresponding to the CDF F_w of the LP, even for $a = 0$ and equal weights.

Table 3.1: Overview of the aggregation methods for probabilistic forecasts, with F_i and Q_i denoting the predictive CDFs and quantile functions of the individual components models. The column 'Parameters' indicates which parameters are estimated based on data, following the procedure described in Section 3.1.2.

Abbr.	Scale	Formula	Parameters	Estimation
LP	Probability	$F_w = \frac{1}{n}\sum_{i=1}^{n} F_i$	-	-
$\mathrm{V}_0^=$	Quantile	$Q_w = \frac{1}{n}\sum_{i=1}^{n} Q_i$	-	-
$\mathrm{V}_a^=$	Quantile	$Q_w = \frac{1}{n}\sum_{i=1}^{n} Q_i + a$	$a \in \mathbb{R}$	CRPS
V_0^w	Quantile	$Q_w = w_0 \sum_{i=1}^{n} Q_i$	$w_0 \geq 0$	CRPS
V_a^w	Quantile	$Q_w = w_0 \sum_{i=1}^{n} Q_i + a$	$w_0 \geq 0$, $a \in \mathbb{R}$	CRPS

choice for estimating the combination parameters. Given NN models estimated based on the training set (where the validation set is used to determine hyperparameters), we estimate the coefficients of the VI approaches separately in a second step based on the validation set, which can be seen as a post-hoc calibration step (Guo et al., 2017). During this second step, the component models with quantile functions Q_i, $i = 1, \ldots, K$, are considered fixed and we only vary the combination parameters in equation (3.2). In the following, we will restrict our attention to fixed training and validation sets, but an extension of the approach described here to a cross-validation (CV) setting is straightforward. Table 3.1 provides an overview of the abbreviations and important characteristics of the different forecast aggregation methods we will consider below.

VI (in the form of $\mathrm{V}_0^=$) has recently received more research interest in the ML literature and has for example been used by Kirkwood et al. (2021) and Kim et al. (2021) to aggregate probabilistic predictions. Related work in the statistical literature includes comparisons to the LP which demonstrate that VI tends to perform better than the LP (Lichtendahl et al., 2013; Busetti, 2017).

Regarding the different NN-based methods for probabilistic forecasting that will be introduced in Section 3.2, we now consider the special case of VI for location-scale families. Given a CDF $F_{(0)}$, a distribution is said to be an element of a *location-scale family* if its CDF F satisfies

$$F(y; \mu, \sigma) = F_{(0)}\left(\frac{y-\mu}{\sigma}\right), \quad y \in \mathbb{R}, \tag{3.3}$$

where $\mu \in \mathbb{R}$ denotes the location and $\sigma > 0$ the scale parameter. Popular examples of location-scale families include the normal and logistic distributions. Unlike the LP, which results in a widespread, multimodal distribution, VI is *shapepreserving* for location-scale families (Thomas and Ross, 1980). Shapepreserving here means that if the individual forecasts are elements of the same location-scale family, the aggregated forecast is as well. Further, the parameters of the aggregated forecast μ^{VI} and σ^{VI} are given by the weighted averages of the

individual parameters μ_i and σ_i, $i = 1, \ldots, K$, together with the intercept a in case of the location parameter, that is,

$$\mu^{\mathrm{VI}} = a + \sum_{i=1}^{K} w_i \mu_i, \quad \text{and} \quad \sigma^{\mathrm{VI}} = \sum_{i=1}^{K} w_i \sigma_i. \tag{3.4}$$

Here, we will only consider the case of $w_i = w_0$ for $i = 1, \ldots, K$. Lichtendahl et al. (2013), who compare the theoretical properties of the LP and $\mathrm{V}_0^{=}$, note that the aggregated predictive distributions both yield the same mean but the VI forecasts are sharper, that is, the VI predictive distribution has a variance smaller or equal to that of the LP.

To highlight the effects of the individual VI parameters, we note that the intercept a only has an effect on the location of the resulting aggregated distribution, while the weight w_0 has an effect on both the location and the spread. If it is larger than 1, the spread increases compared to the average spread of the individual forecasts, and it decreases for values smaller than 1. However, a weight not equal to 1 also shifts the location of the distribution. Figure 3.1 illustrates this in the exemplary case of two normal distributions.

3.2 Aggregating exemplary types of forecast distributions

In the context of weather prediction, we propose a framework for NN-based probabilistic forecasting in Section 4.3 that encompasses different approaches to obtain distribution forecasts as the output of an NN and forms the basis of our work here. However, we will leave the introduction of this framework that includes three network variants to Section 4.3. Here, we will consider only the associated types of forecast distribution, which coincide with those presented in Section 2.3.

While the three network variants differ in their characterization of the forecast distribution, their use in practice shares a common methodological feature that constitutes the main motivation for our work here. As discussed at begin of the section, extant practice in NN-based forecasting often relies on an ensemble of NN models trained based on randomly initialized weights and batches to account for the randomness of the stochastic gradient descent methods applied in the training process. This raises the question of how the three types of distribution forecasts can be combined using the aggregation methods described in Section 3.1, which we will discuss below.

3.2.1 Parametric forecast distribution

In the *distributional regression network* (DRN) approach, the forecasts are issued in the form of a parametric distribution. Under the parametric assumption F_θ, the predictive distribution is characterized by the distribution parameter (vector) $\theta \in \Theta \subset \mathbb{R}^d$, where Θ is the parameter space. Different variants of the DRN approach have been proposed over the past years and

can be traced back to at least Bishop (1994). While Lakshminarayanan et al. (2017) and Rasp and Lerch (2018) use a normal distribution with $\theta = (\mu, \sigma)$, where $\mu \in \mathbb{R}$ is the mean and $\sigma > 0$ the standard deviation, Bishop (1994) and D'Isanto and Polsterer (2018) use a mixture of normal distributions. In the case studies on probabilistic wind gust forecasting (Sections 3.4 and 5.3), we use a zero-truncated logistic distribution with $\theta = (\mu, \sigma)$, where $\mu \in \mathbb{R}$ is the location and $\sigma > 0$ the scale parameter.[4] Extensions of the DRN approach to other parametric families are generally straightforward provided that analytical closed-form expressions of the selected loss function are available (e.g., Ghazvinian et al., 2021; Chapman et al., 2022).

Both Lakshminarayanan et al. (2017) and Rasp and Lerch (2018) generate an ensemble of networks based on random initialization. While Lakshminarayanan et al. (2017) propose to use the LP to aggregate the forecast distributions, Rasp and Lerch (2018) instead combine the forecasts by averaging the distribution parameters. Since the normal distribution (which we will also employ in the simulation study below) is a location-scale family, parameter averaging is equivalent to $\mathrm{V}_0^{=}$. Although the logistic distribution also forms a location-scale family, the truncated variant does not, and parameter averaging is not equivalent to $\mathrm{V}_0^{=}$. However, in the context of the case study, we found the differences between parameter averaging and $\mathrm{V}_0^{=}$ to be negligibly small and, in this particular case, therefore approximated the VI approaches by the corresponding parameter averages. To evaluate the LP forecasts, we draw a random sample of size 1,000 from the mixture distribution by first randomly choosing an ensemble member and then generating a random draw from the corresponding distribution.

3.2.2 Bernstein quantile function

Bremnes (2020) proposes a semiparametric extension of the DRN approach we refer to as *Bernstein quantile network* (BQN), where the probabilistic forecast is given by the Bernstein quantile function in equation (2.33). To aggregate ensembles of BQN forecasts, Bremnes (2020) average the individual basis coefficient values across ensemble members. This is equivalent to $\mathrm{V}_0^{=}$, which can be seen by plugging in Bernstein quantile functions (equation (2.33)) in the VI quantile function in equation (3.2),

$$\begin{aligned} Q_w(\tau) &= a + \sum_{i=1}^{K} w_i \left(\sum_{l=0}^{d} \alpha_{il} B_{ld}(\tau) \right) \\ &= a + \sum_{l=0}^{d} \left(\sum_{i=1}^{K} w_i \alpha_{il} \right) B_{ld}(\tau), \quad \tau \in [0,1]\,, \end{aligned} \tag{3.5}$$

where α_{il} is the coefficient of the lth basis polynomial of the ith ensemble member, $i = 1, \ldots, K$, $l = 0, \ldots, d$.

[4] In the context of weather prediction, the choice of a (parametric) forecast distribution is discussed in Section 4.2.1.

Since a closed form of the CDF or density of a BQN forecast is not readily available, the LP cannot be expressed in a similar fashion. Analogous to DRN, the evaluation of the LP forecasts will therefore be based on a random sample of size 1,000 drawn from the aggregated distribution. Here, the inversion method allows to sample from the individual BQN forecasts. Further, the VI forecasts are evaluated based on a sample of 100 equidistant quantiles.[5]

3.2.3 Piecewise uniform distribution

The last method considered here is the *histogram estimation network* (HEN), which is based on the piecewise uniform distribution presented in Section 2.3.3. Variants of this approach have been proposed in a variety of applications (e.g., Gasthaus et al., 2019; Li et al., 2021). Here, we consider the case of fixed bins, i.e., the output of the network is a set of bin probabilities.

Regarding the aggregation of an ensemble of piecewise uniform distributions with fixed bins, the LP is equivalent to averaging the bin probabilities. To see this, we apply the LP (equation (3.1)) to CDFs of a piecewise uniform distribution (equation (2.41)), that is,

$$\begin{aligned} F_w(y) &= \sum_{i=1}^{K} w_i \left[\sum_{l=1}^{N} \left(p_{il} \frac{(y|_{I_l}) - b_{l-1}}{b_l - b_{l-1}} \cdot \mathbb{1}\{b_{l-1} \leq y\} \right) \right] \\ &= \sum_{l=1}^{N} \left[\left(\sum_{i=1}^{K} w_i p_{il} \right) \frac{(y|_{I_l}) - b_{l-1}}{b_l - b_{l-1}} \cdot \mathbb{1}\{b_l \leq y\} \right], \end{aligned} \tag{3.6}$$

where $y \in \mathbb{R}$ and p_{il} is the probability of the lth bin for the ith ensemble member, $i = 1, \ldots, K$, $l = 1, \ldots, N$. An exemplary application of the LP for an approach akin to HEN forecasts in a stacked NN can be found in Clare et al. (2021).

By contrast to the LP, the VI approach exhibits a particular advantage for HEN forecasts in that it results in a finer binning than the individual HEN models. To illustrate this effect, we recall that the quantile function of a piecewise uniform distribution is a piecewise linear function with edges depending on the accumulated bin probabilities (equations (2.44)–(2.46)). Therefore, the resulting VI quantile function is a piecewise linear function with one edge for each accumulated probability of the individual forecasts. As the forecast probabilities differ for each member of the deep ensemble, the associated quantile functions are subject to a different binning. Since the set of edges of the aggregated VI forecast is given by the union of all individual edges, this leads to a smoothed final forecast distribution with a finer binning than the individual model runs that differs for every forecast case, and eliminates the potential downside of too coarse fixed bin edges. Figure 3.2 illustrates the effects of the LP and $\mathrm{V}_0^{=}$ for two exemplary piecewise uniform distributions. In the simulations study, the edges are given by 50 equidistant empirical quantiles of the training data (unique to the second digit), and

[5] The numbers of samples and quantiles were chosen based on simulation experiments and theoretical considerations. Compared to random samples from the forecast distributions, a smaller number of equidistant quantiles is required to achieve approximations of the same accuracy, see Krüger et al. (2021) and references therein for a discussion of sample-based estimation of the CRPS.

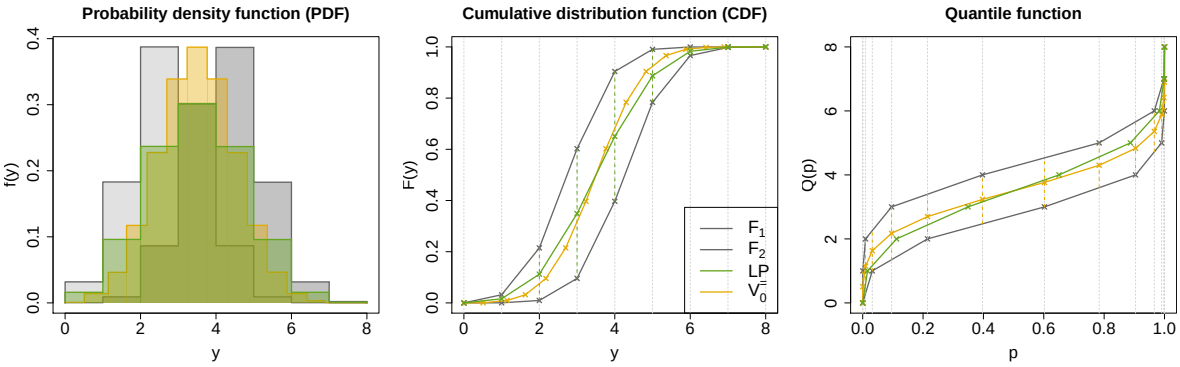

Figure 3.2: PDF, CDF and quantile function of two piecewise uniform distributions F_1 and F_2 together with forecasts aggregated via the LP and $\mathrm{V}_0^=$. The dashed vertical lines indicate the binning with respect to F_1, F_2 and F_w for the CDF plot and with respect to Q_w in the quantile function plot.

for the case study, we use a semiautomated procedure specific to the application, which is described in detail in Section 5.3.2.

3.3 Simulation study

We compare the performance of the five aggregation methods for each of the three network variants in a simulation study. The simulation setting is adopted from Li et al. (2021), who investigate a variant of the HEN approach. From the six models they propose for the data-generating process, we consider two and skip results for the remaining four, as they provide no further insights.

We do not tune the specific architectures and hyperparameters of the individual NN models in each of the scenarios of the simulation study, but instead use the configurations described in Section 5.3.2 that have proven to work well in the corresponding application. This is done intentionally in order to also generate forecasts that are not well-calibrated or subject to systematic biases, which allows us to assess the performance of the aggregation methods in situations when the forecasts are not optimal, see, in particular, the results for Scenario 2 reported below.

For each scenario, we generate training sets of size 6,000 and test sets of size 10,000. The simulations are repeated 50 times. We generate a series of 40 individual network ensemble members for each of the three network variants and consider aggregation of the first K members, where $K \in \{2, 4, \ldots, 40\}$. As benchmark, we will consider an optimal probabilistic forecast based on the inherent uncertainty of the data-generating process denoted by the noise term ϵ.[6] In the following, the *continuous ranked probability skill score* (CRPSS) will be calculated via equation (2.5) using the average CRPS of the individual networks for $\overline{S}_n^{F_0}$ and

[6] Note that the simulations are based on a finite sample, so even the optimal forecast might result in a small bias or an empirical coverage not equal to the nominal value.

the CRPS of the optimal forecast for $\overline{S}_n^{F^*}$ with n being the size of the underlying sample.[7]

Scenario 1

As our first simulation scenario, we consider a linear model with normally distributed errors. Based on a random vector of predictors $\boldsymbol{X} \in \mathbb{R}^5$, which serves as the input of the networks, and the random coefficient vectors $\boldsymbol{\beta_1}, \boldsymbol{\beta_2} \in \mathbb{R}^5$, which are fixed for each run of the simulation and unknown to the forecaster, the target variable Y is calculated via

$$Y = \boldsymbol{X}^T \boldsymbol{\beta_1} + \epsilon \cdot \exp\left(\boldsymbol{X}^T \boldsymbol{\beta_2}\right), \tag{3.7}$$

where $\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I_5})$, $\boldsymbol{\beta_1} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I_5})$, $\boldsymbol{\beta_2} \sim \mathcal{N}(\mathbf{0}, 0.45^2 \boldsymbol{I_5})$ and $\epsilon \sim \mathcal{N}(0, 1)$. The optimal forecast F^* is then given by

$$Y \mid \boldsymbol{X}, \boldsymbol{\beta_1}, \boldsymbol{\beta_2} \sim \mathcal{N}\left(\boldsymbol{X}^T \boldsymbol{\beta_1}, \exp\left(2\boldsymbol{X}^T \boldsymbol{\beta_2}\right)\right) = F^*. \tag{3.8}$$

The key results for this simulation scenario are summarized in Figure 3.3, which shows different evaluation metrics averaged over the 50 repetitions of the simulation study. We start by comparing the aggregation methods for DRN, where the CRPSS indicates that aggregation via the VI approaches improves the network average by up to 12.5%, while the LP improves the forecast performance by at most 2.5%. Here, the best VI approach is to fix the intercept and weights instead of estimating them from the training data. In Figure 3.3, the *relative weight difference* of the estimated weight w_0 and a standardized weight for an ensemble of size K, given by $\delta_K(w_0) := Kw_0 - 1$, illustrates that the estimated weights are not equal to standardized weights. The flat PIT histograms in Figure 3.4 indicate that the individual component forecasts are already well-calibrated and corrections via coefficient estimation are not necessary. The average PI length of the network forecasts, which is identical to that of $\mathrm{V}_0^=$ and $\mathrm{V}_a^=$, is smaller than that of the optimal forecast. Note that having sharper forecasts than F^* comes at the cost of a lack in calibration. Comparing the aggregation methods, we find that the LP increases the PI length as expected due to its theoretical properties. V_0^w and V_a^w here increase the PI length because their estimated weights are larger than standardized ones. All aggregation methods increase the empirical coverage, which improves the predictive performance, because the coverage of the network average is smaller than the nominal value. In terms of accuracy, all methods are unbiased, since they are approximately as accurate as the optimal forecast.

For BQN, the results are qualitatively similar, however, since the BQN forecasts are not as well-calibrated as those of DRN, there are some differences which we highlight in the following. The estimated weights of V_0^w and V_a^w are larger than standardized ones and result

[7] Note that this does not correspond to the mean improvement over the individual forecasts. However, averaging the median skill scores of the individual ensemble member predictions over the repetitions of the simulations yields qualitatively analogous results (not shown).

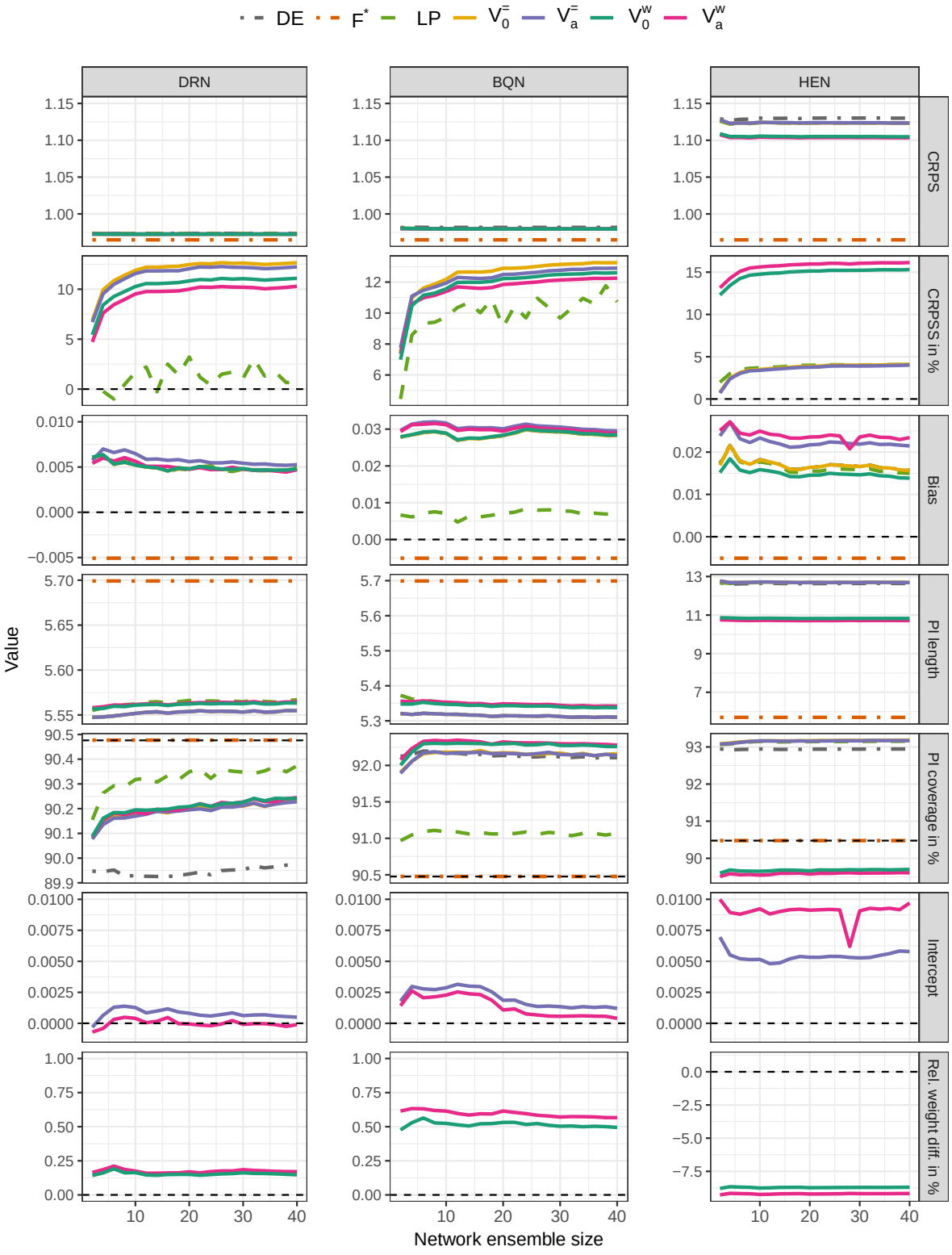

Figure 3.3: Different evaluation metrics, the estimated intercept a and the relative weight difference δ_n for the three network variants in Scenario 1 of the simulation study, where DE denotes the average score of the deep ensemble members. Note the different scales on the y-axis.

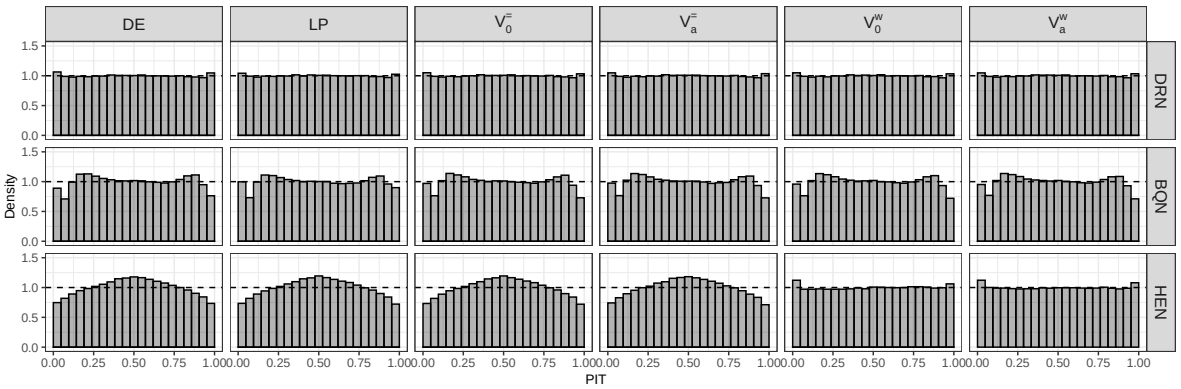

Figure 3.4: PIT histograms for the three network variants of the deep ensemble (DE) and aggregated forecasts for an ensemble of size 2 in Scenario 1.

in a smaller CRPSS difference to $V_0^=$, which still performs best. The V_0^w and V_a^w forecasts are therefore less sharp than the network average and as sharp as the LP. The empirical coverage of the individual BQN forecasts is larger than the nominal value, thus that of the forecasts aggregated via VI approaches is as well. Interestingly, the LP decreases the coverage and is closest to the nominal value. Further, the VI forecasts are positively biased, the LP is instead close to being unbiased. Although the LP performs favorable in terms of the empirical coverage and accuracy, it performs worse than the VI approaches, even though the difference is smaller than in the case of DRN.

In contrast to DRN and BQN, the HEN forecasts are not well-calibrated but instead overdispersed, as indicated by the bulk-shaped histograms in Figure 3.4. In addition to the lack of calibration, the forecasts are also not sharp since the PIs are more than twice as large as those of the optimal forecast. These deficiencies result in a substantially worse CRPS compared to DRN and BQN. While the LP, $V_0^=$ and $V_a^=$ are unable to correct the systematic miscalibration, V_0^w and V_a^w result in well-calibrated forecasts, which is indicated by the flat PIT histograms in Figure 3.4. The estimated weights are smaller than standardized ones for all ensemble sizes, therefore the forecasts become sharper. The PI coverage of the overdispersed forecasts is, as expected, 2.5% larger than the nominal value. The corrections of V_0^w and V_a^w result in coverages closer to and even smaller than the nominal value. Further, note that V_a^w estimates smaller weights than V_0^w, but also a positive intercept larger than that of $V_a^=$ in order to balance the effect of the weights on the location of the aggregated distribution. However, the positive intercepts estimated by $V_a^=$ and V_a^w result in a larger bias. The correction of the overdispersion is also reflected in the CRPSS, where V_a^w and then V_0^w outperform the other approaches by a wide margin. Note that the LP improves the predictive performance and performs equally well as $V_0^=$ and $V_a^=$. However, although all aggregation methods correct systematic errors and improve predictive performance, the aggregated forecasts are still not competitive to those of DRN and BQN in terms of the CRPS.

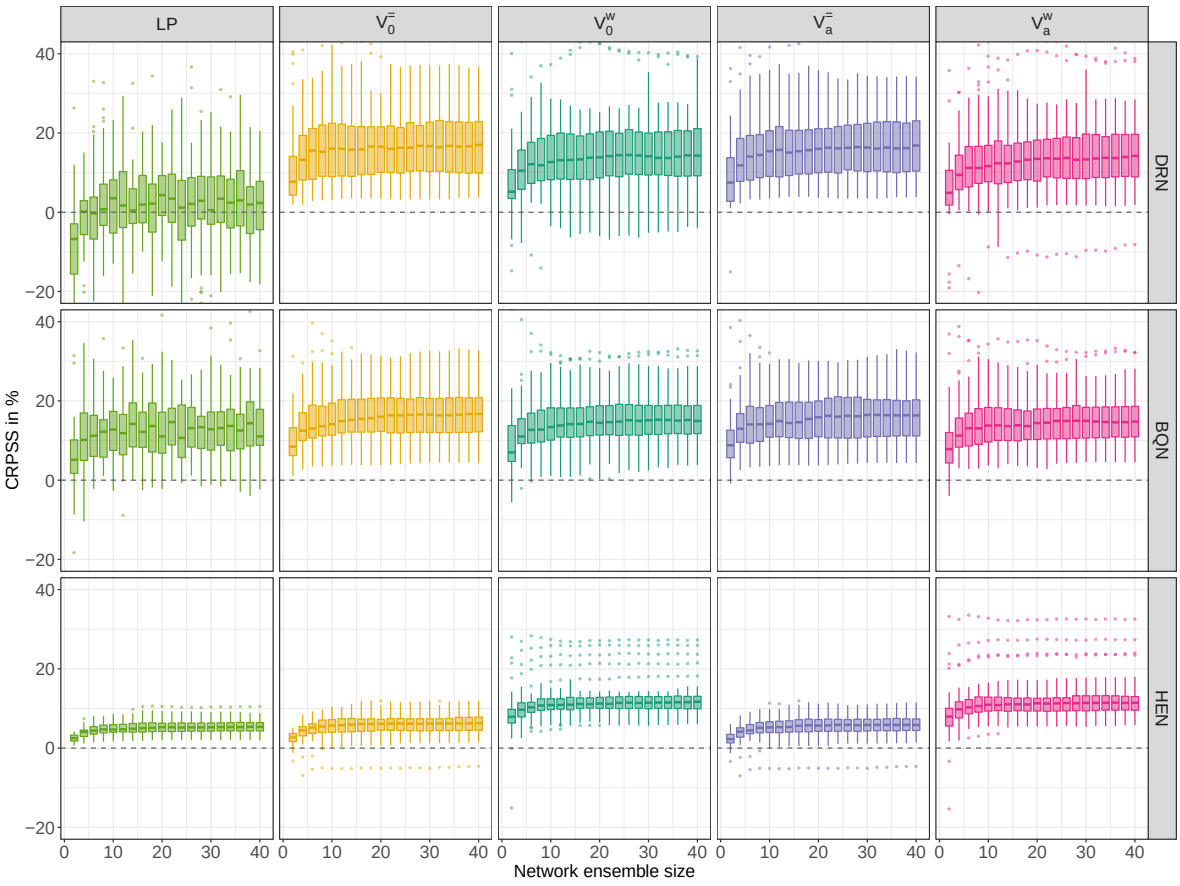

Figure 3.5: Boxplots over the CRPSS values of the 50 runs in Scenario 1 of the simulation study. Note some outliers and that the boxes of the LP for BQN and DRN are cut off to improve readability.

Finally, we investigate the effect of the ensemble size on the performance of the aggregation methods, in particular on the CRPSS, considering the variability over the 50 runs (Figure 3.5). For all network variants and aggregation methods, most improvement is obtained up to ensembles of size 10. The median CRPSS increases up to a size of 20, after which only minor further improvements can be observed. Interestingly, the variability over the runs does not decrease for larger ensemble sizes. Comparing the aggregation methods, more outliers are observed for methods that estimate parameters. For DRN, we see that parameter estimation may result in forecasts worse than the network average in a few cases, on the other hand the same results are obtained for $V_0^{=}$ forecasts in case of HEN. Regarding systematic differences between the network variants, we find that the variation across simulation runs is notably lower for HEN. This can partly be explained by the fact that the DRN and BQN forecasts are much closer to the optimal forecasts and thus small absolute deviations in the CRPS result in larger differences in the skill.

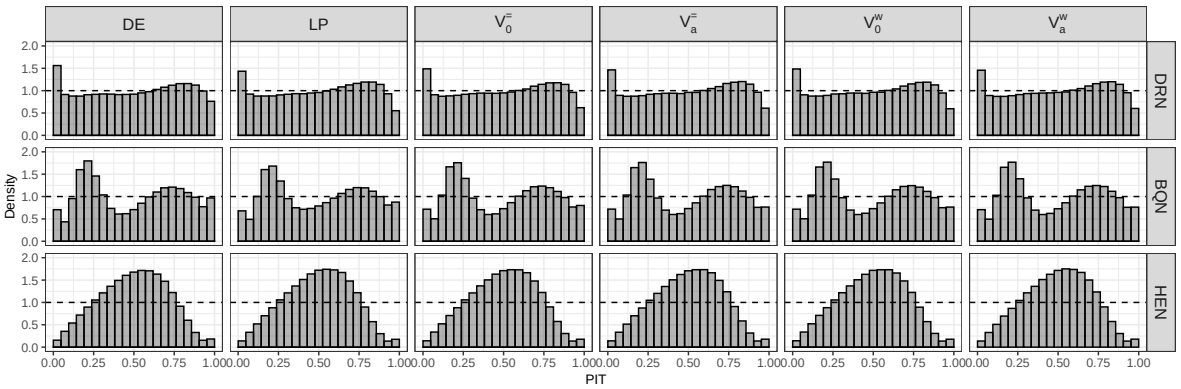

Figure 3.6: PIT histograms for the three network variants of the deep ensemble (DE) and aggregated forecasts for an ensemble of size 2 in Scenario 2.

Scenario 2

In the second scenario, we consider a skewed distribution with a nonlinear mean function. The target variable Y is defined by

$$Y = 10 \sin(2\pi X_1 X_2) + 20 (X_3 - 0.5)^2 + 10 X_4 + 5 X_5 + \epsilon, \tag{3.9}$$

where $\boldsymbol{X} = (X_1, \ldots, X_5)^T$, $X_1, \ldots, X_5 \overset{iid}{\sim} \mathcal{U}(0,1)$, and $\epsilon \sim \text{SkewNormal}(0, 1, -5)$. The optimal forecast is given by the conditional distribution of $Y \mid \boldsymbol{X}$, that is,

$$F^* = \text{SkewNormal}\Big(10 \sin(2\pi X_1 X_2) + 20 (X_3 - 0.5)^2 + 10 X_4 + 5 X_5, 1, -5\Big). \tag{3.10}$$

PIT histograms of the individual and aggregated forecasts for the different network variants are shown in Figure 3.6. In contrast to the first scenario, none of the network variants produces calibrated forecasts and their PIT histograms indicate systematic deviations from uniformity. As to be expected due to the wrong distributional assumption, DRN based on a normal distribution is not able to yield calibrated forecasts for an underlying skewed normal distribution, but also the semiparametric BQN and HEN forecasts fail to provide calibrated forecasts. The HEN forecasts are strongly overdispersed and again result in the worst CRPS among the network variants (Figure 3.7). None of the aggregation methods is able to correct the systematic lack of calibration for all of the network variants. That said, aggregation still improves the overall predictive performance in terms of the CRPS.

For DRN, we find that the LP outperforms the VI approaches. Interestingly, this is the case even though the LP forecasts are the least sharp and have a higher PI coverage that is farther away from the nominal value. In contrast to the first scenario, even the aggregated DRN forecasts perform notably worse than BQN. The VI approaches perform equally well and increase the coverage of the forecasts such that they are closer to the nominal value. While $V_a^=$ and V_0^w estimate coefficients close to the nominal values, V_a^w estimates larger weights,

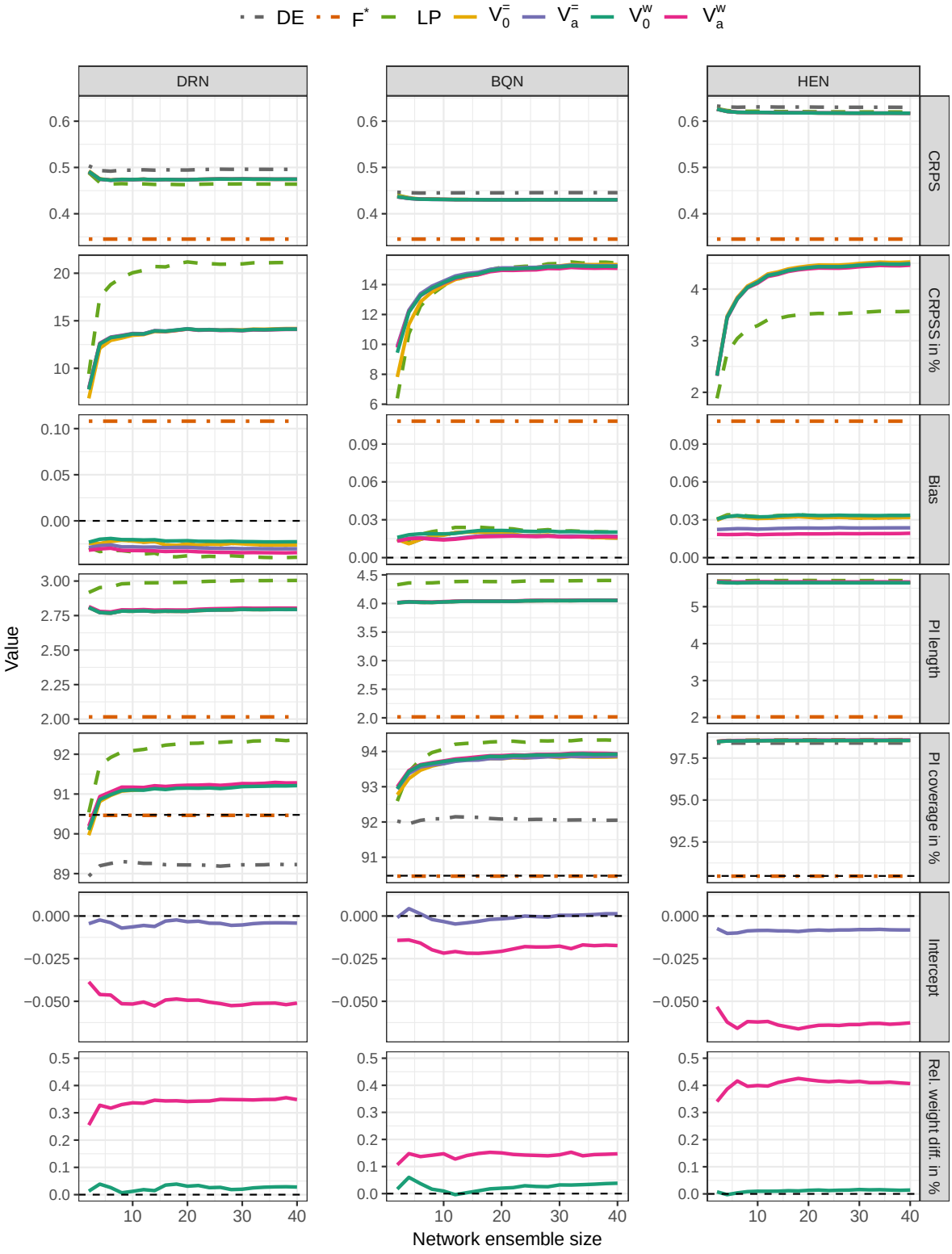

Figure 3.7: Evaluation metrics for the three network variants in Scenario 2, where DE denotes the average score of the deep ensemble members. Mind the different scales on the y-axis.

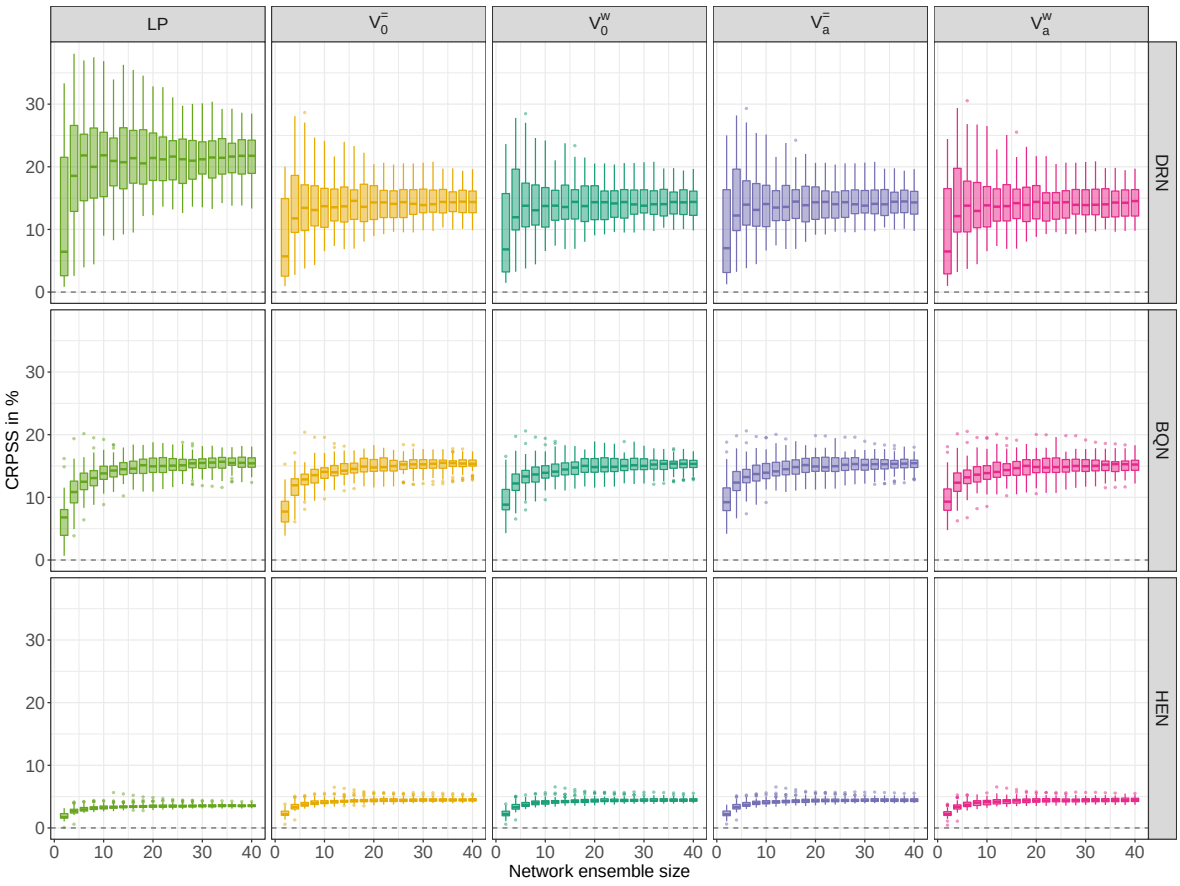

Figure 3.8: Boxplots over the CRPSS values of the 50 runs in Scenario 2 of the simulation study.

and therefore yields larger PIs, and a negative intercept in order to balance the shift in the location. Still, V_a^w does not outperform the other VI approaches.

The results are again qualitatively similar for BQN. The main difference is that the LP does not outperform the VI approaches, as all aggregation methods result in an improvement in terms of the mean CRPS of up to 16%. Further, the LP again yields the least sharp forecasts and all methods increase the PI coverage.

Next, we consider the HEN forecasts. In contrast to Scenario 1, weight estimation via V_0^w and V_a^w is not able correct the systematic errors and outperform the other approaches. All VI approaches perform equally well and outperform the LP, which still improves the network average. The LP yields the least sharp forecasts, followed by V_a^w that estimates weights larger than standardized ones together with a negative intercept, as for DRN and BQN. The negative intercepts of $V_a^=$ and V_a^w improve the accuracy, as they decrease the forecast bias.

Regarding the effect of the ensemble size, the largest improvements of the aggregation methods are again obtained for up to 10 ensemble members and only minor improvements can be observed for sizes larger than 20 (Figure 3.8). In contrast to Scenario 1, it can be noted that the variability over the runs decreases as the ensemble size increases, and that

the degree of variability is similar for all aggregation methods within one network variant. A direct comparison of the network variants indicates that the variability generally increases with the overall skill of the aggregated forecast.

3.4 Case study

Our case study focuses on the application of the aggregation methods to probabilistic wind gust forecasting over Germany using forecast distributions obtained as the output of NN methods. For context on the underlying problem, we refer the reader to Section 4.1, while the case study on probabilistic wind gust forecasting including a description of the data, the configuration of the NN models and a comparison to other probabilistic forecasting methods can be found in Section 5.3.

Here, we focus on different subsets of the data, namely, we consider 4 of the 22 forecast horizons (0, 6, 12, 18 hours), which are referred to as lead times in the context of weather predictions. In the following, we will typically evaluate the predictive performance aggregated over those lead times and note that while there are minor differences across lead times, the results are qualitatively similar and all the main conclusions are valid for all the considered lead times.

For each of the lead times, we generate an ensemble of 100 models for each of the network variants based on random initialization, which form the basis for our study of the different aggregation methods. We randomly draw a subset of these 100 models for each of the considered ensemble sizes $K \in \{2, 4, \ldots, 40\}$ and repeat this procedure 20 times to account for uncertainties. Therefore, 20 aggregated forecasts based on a pool of 100 network ensemble members are generated for each model variant and each ensemble size.

Note that the underlying distribution of the target variable is of course unknown in the case study, and following common practice the observed value is used as the (hypothetical) optimal forecast resulting in a CRPS of 0. Hence, we calculate the CRPSS via equation (2.6). The magnitude of the CRPSS values of the simulation and case study is thus not directly comparable.

Figure 3.9 summarizes the key results of the case study. Applying the aggregation methods to the DRN and BQN forecasts leads to similar results as in the simulation study. Although the LP improves predictive performance with a skill of up to 1.6%, the VI approaches are superior to the LP. Among the considered VI approaches, coefficient estimation leads to better predictions with V_a^w and $\mathrm{V}_a^=$ performing best, followed by V_0^w and $\mathrm{V}_0^=$.

As expected, the LP yields less sharp forecasts than the network average indicated by larger PIs, which are the least sharp for DRN. V_a^w also issues less sharp forecasts than the network average, which is identical to that of $\mathrm{V}_0^=$ and $\mathrm{V}_a^=$, and V_0^w produces the sharpest forecasts. This is due to the fact that V_0^w estimates weights smaller and V_a^w larger than the nominal value of $\mathrm{V}_0^=$. As in the simulation study, V_a^w estimates a more extreme intercept

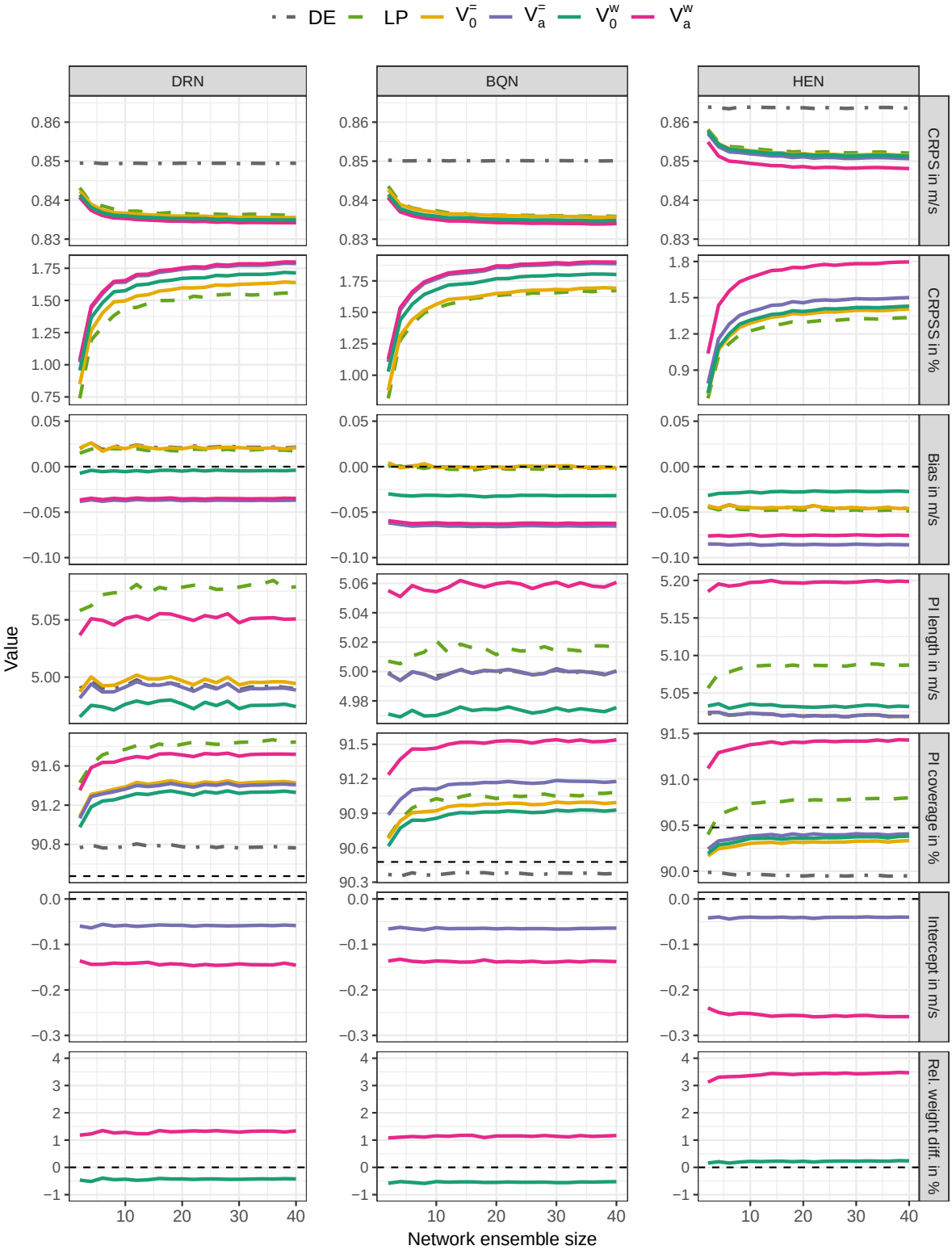

Figure 3.9: Evaluation metrics for the three network variants aggregated over all lead times considered in the case study, where DE denotes the average score of the deep ensemble members. Note the different scales on the vertical axis.

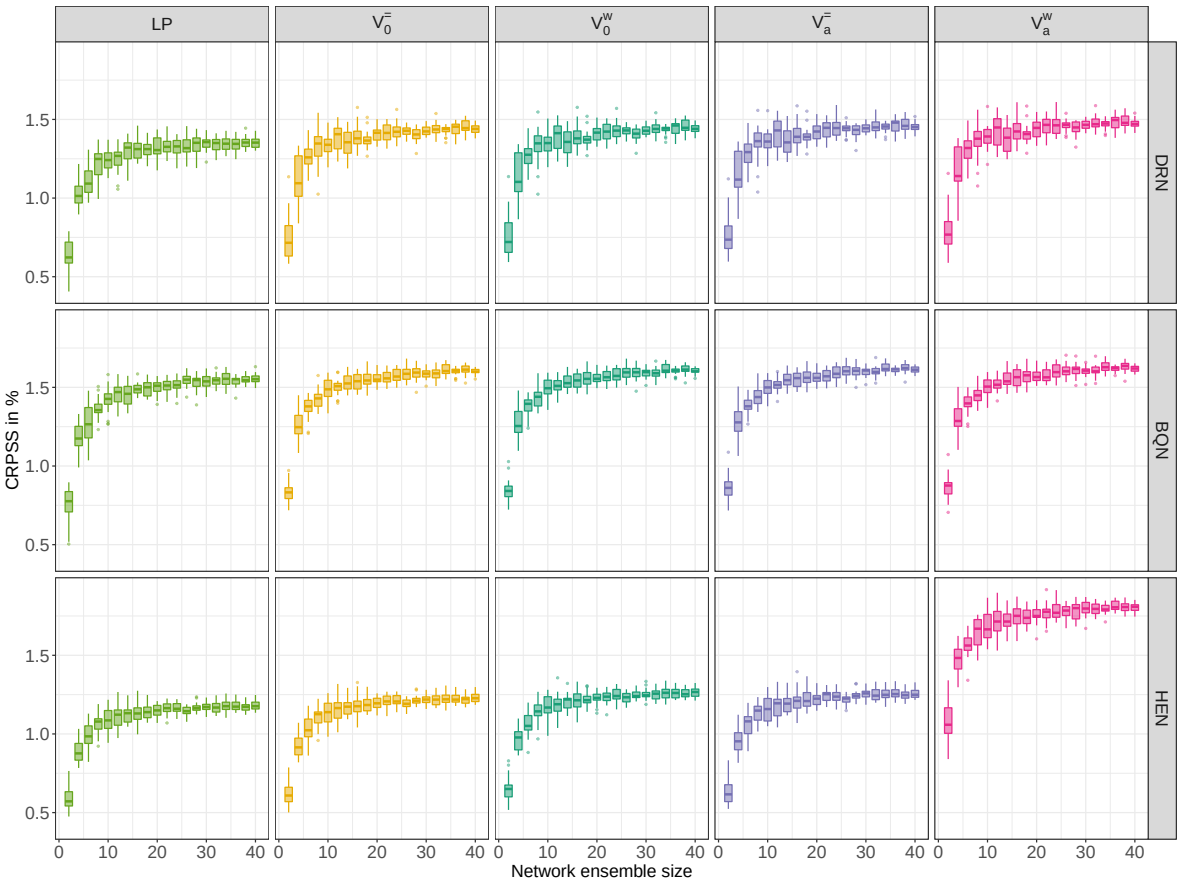

Figure 3.10: Boxplots over the CRPSS values of the 20 draws for each ensemble size in the case study for a lead time of 18 hours.

than $V_a^=$, which balances the effect of the weight estimation. The PI coverage increases for all aggregation methods and both network variants. For both network variants, the V_0^w forecasts have the smallest coverage closest to the nominal value, whereas V_a^w results in a coverage larger than the other VI approaches. Only for DRN, the LP has a larger PI coverage.

The results of the aggregated HEN forecasts are again qualitatively different from those of DRN and BQN, as the HEN method does not perform as well as the other two methods and is subject to more systematic errors. Although the ranking of the aggregation methods is identical to that of DRN and BQN, the magnitude of the differences in the CRPSS for the superior method is notably larger, and since V_a^w is able to improve some of the systematic errors, it clearly outperforms the other approaches. The most significant difference to the other aggregation methods is that V_a^w estimates more extreme coefficients. As for BQN, this results in the largest PIs and largest coverage, in both cases followed by the LP.

Regarding the accuracy of the forecasts produced by the different aggregation methods, the results are qualitatively similar for all three network variants. The two methods that estimate an intercept have the largest absolute biases. This is a somewhat counterintuitive result, since it can be expected that including an intercept should enable the correction of systematic

biases. As noted in Section 5.3, there are minor structural differences in the distribution of the observed values in the test and validation datasets. Due to the average observed values in the test data being somewhat smaller than those in the validation dataset, the data the coefficients are estimated on is not fully representative of the test data.

To assess the effect of the ensemble size on the predictive performance in Figure 3.10, we pick one specific lead time, namely 18 hours, to avoid distortions in the boxplots caused by the minor variations in the magnitude of the improvements over lead times. The results coincide with the corresponding main conclusions of the simulation study in that we observe most improvement up to ensembles of size 10 and only minor for ensembles of size larger than 20. In the case study, the improvement up to size 10 is more pronounced than in the simulation study and strongly suggests that a network ensemble should include at least 10 members. Finally, we note that the variability of the CRPSS decreases for larger ensemble sizes.

3.5 Discussion and conclusions

We have conducted a systematic comparison of aggregation methods for the combination of distribution forecasts from ensembles of NNs based on random initialization, so-called deep ensembles. In doing so, this section aims to reconcile and consolidate findings from the statistical literature on forecast combination and the ML literature on ensemble methods. Specifically, we propose a general VI framework where quantile functions of the forecast distributions can be flexibly combined, and compare to the results of the widely used LP, where the probabilistic forecasts are linearly combined on the scale of probabilities. For deep ensembles of three variants of NN-based models for probabilistic forecasting that differ in the characterization of the output distribution, aggregation with both the LP and VI improves the predictive performance. The VI approaches show superior performance compared to the LP. For example, given ensemble members that are already calibrated, $\mathrm{V}_0^{=}$ preserves the calibration and improves the predictive accuracy while the LP decreases sharpness with more dispersed forecasts. If the individual forecast distributions are subject to systematic errors such as biases and dispersion errors, coefficient estimation via $\mathrm{V}_a^{=}$, V_0^{w} and V_a^{w} is able to correct these errors and improve the predictive performance considerably, otherwise $\mathrm{V}_0^{=}$ should be preferred. While these combination approaches require the estimation of additional combination coefficients, the computational costs are negligible compared to the generation of the NN-based probabilistic forecasts and can be performed on the validation data without restricting the estimation of the NNs.

Even though forecast combination generally improves the predictive performance, Scenario 2 of the simulation study demonstrates that for example the lack of calibration of the severely misspecified individual forecast distributions cannot be corrected by the aggregation methods considered here. In the context of NNs and deep ensembles, the calibration of (ensemble) predictions and recalibration procedures have been a focus of much recent research interest

(Guo et al., 2017; Ovadia et al., 2019). For example, in line with the results of Gneiting and Ranjan (2013), deep ensemble predictions based on the LP were found to be miscalibrated and should be recalibrated after the aggregation step (Rahaman and Thiery, 2021; Wu and Gales, 2021). A wide range of recalibration methods, which simultaneously aggregate and calibrate the ensemble predictions (such as the $V_a^=$, V_0^w and V_a^w approaches presented in Section 3.1.2 for VI), have been proposed in order to correct the systematic errors introduced by the LP in the context of probability forecasting for binary events (Allard et al., 2012). For example, the beta-transformed LP composites the CDF of a beta distribution with the LP (Ranjan and Gneiting, 2010), and Satopää et al. (2014) propose to aggregate probabilities on a log-odds scale. Some of these approaches can be readily extended to the case of forecast distributions (Gneiting and Ranjan, 2013). For VI, more sophisticated approaches that allow the weights to depend on the quantile levels might improve the predictive performance (Kim et al., 2021). Further, moving from a linear combination function towards more complex transformations allowing for nonlinearity might help to correct more involved calibration errors.

We have restricted our attention to ensembles of NN-based probabilistic forecasts generated based on random initialization. While such deep ensembles have been demonstrated to work well in many settings (Lee et al., 2015; Fort et al., 2019; Ovadia et al., 2019), a variety of alternative approaches for uncertainty estimation in NNs has been proposed including Bayesian NNs (Neal, 2012) or generative models (Mohamed and Lakshminarayanan, 2016). A particularly prominent approach to deal with the uncertainty in the estimation of NNs is dropout (Srivastava et al., 2014; Gal and Ghahramani, 2016). Dropout can not only be used as a regularization method during estimation but also for prediction, which results in an ensemble of forecasts and is readily applicable for the different variants of NN methods considered here. Compared to deep ensembles based on random initialization, a potential advantage of dropout-based ensembles is that the lower computational costs make the generation of larger ensembles more feasible. The aggregation methods we investigated are agnostic to the generation of the ensembles provided that they can be considered as realizations of the same basic type of model, and are thus readily applicable to dropout-based ensembles.[8] Therefore, an interesting avenue for future work is to investigate the performance of the combination methods for different approaches to generate NN-based probabilistic forecasts, e.g., within the framework of comprehensive simulation testbeds (Osband et al., 2021).

Concerning the network variants considered in this chapter, averaging the output parameters is equivalent to $V_0^=$ for BQN, and DRN based on a location-scale family. However, in case of HEN, averaging the bin probabilities, which are the output of the NN, is equivalent to the LP, while quantile averaging results in a refinement of the binning underlying the forecast distribution. This property offers an interesting pathway for future work, as deep ensembles

[8] In experiments with dropout ensembles in the context of the case study (not shown), we found that aggregating forecast distributions improves the predictions, but the overall performance of both the individual and combined dropout-based forecasts is substantially worse compared to the ensembles considered here.

of classification NNs for ordinal target variables can be used to approximate a continuous forecast based on piecewise uniform distributions. This might be of particular interest in cases, where the target variable is observed only in ordinal classes, but a continuous forecast is desired. First, we define a fixed binning that allows to identify a set of class probabilities with a piecewise uniform distribution, ideally the underlying class definition already provides a binning. Therefore, we can create an ensemble of piecewise uniform distributions based on a classification NN.[9] Now, aggregating this ensemble of distribution forecasts with VI, we obtain a much finer binning. Using dropout, we can efficiently generate large ensembles that result in a fine binning that approximates the desired continuous distribution.

Finally, we summarize three key recommendations for aggregating distribution forecasts from deep ensembles based on our results:

- To optimize the final predictive performance of the aggregated forecast, the individual component forecasts should be optimized as much as possible.[10] While forecast combination improves predictive performance, it generally did not effect the ranking of the different NN-variants for generating probabilistic forecasts, and is unable to fix substantial systematic errors.

- Generating an ensemble with a size of a least 10 appears to be a sensible choice, with only minor improvements being observed for more than 20 members. This corresponds to the results in Fort et al. (2019) and ensemble sizes typically chosen in the literature (Lakshminarayanan et al., 2017; Rasp and Lerch, 2018), but the benefits of generating more ensemble members need to be balanced against the computational costs, and sometimes smaller ensembles have been suggested (Ovadia et al., 2019; Abe et al., 2022).

- Aggregating forecast distributions via VI is often superior to the LP. Thereby, the choice of the specific variant within the general framework depends on potential misspecifications of the individual component distributions, as discussed above.

Note that these conclusions, in particular the superiority of the quantile aggregation approaches, refer to the specific situation of deep ensembles considered here. The property of shapepreservation justifies the use of VI from a theoretical perspective in a setting where the ensemble members are based on the same model and data. If the ensemble members differ in terms of the model used to generate the forecast distribution or the input data they are based on, shapepreservation might not be desired. Instead, a model selection approach based on the LP, which allows for obtaining a multimodal forecast distribution, might better represent the possible scenarios that may materialize.

[9] This does not only hold for NNs, but instead for any ensemble of classification models.

[10] Abe et al. (2022) find that deep ensembles do not offer benefits compared to single larger (that is, more complex) NNs. Our results do not contradict their findings, since we address a conceptually different question and argue that given the generation of a deep ensemble, the individual members' forecasts should be optimized as much as possible. In this situation, a single NN will generally not be able to match the predictive performance of the associated deep ensemble.

CHAPTER 4

Statistical Postprocessing: Methods

Due to the chaotic nature of the atmosphere, weather prediction is a prime example of statistical forecasting. The meteorological community has acknowledged the need to quantify the uncertainty associated with weather prediction and attention has been shifting towards probabilistic forecasting with the first probabilistic prediction systems becoming operational at the *European Centre for Medium-Range Weather Forecasts* (ECMWF) and the *US National Meteorological Center* in 1992 (Toth and Kalnay, 1993; Molteni et al., 1996). Nowadays, these so-called *ensemble forecasts* are key components of operational systems at many national and international weather services. Despite substantial improvements over the past decades (Bauer et al., 2015), ensemble forecasts continue to exhibit systematic errors that require *statistical postprocessing* to achieve accurate and reliable probabilistic forecasts, where statistical postprocessing comprises techniques that learn to correct theses errors based on past forecasts and observations. Statistical postprocessing has therefore become an integral part of weather forecasting and standard practice in research and operations.

After giving a brief description of the generation of weather forecasts and the need for statistical postprocessing, we present a wide range of methods for statistical postprocessing ranging from classic techniques established at meteorological weather services to novel approaches based on modern ML. While Chapter 4 gives a theoretical overview of the postprocessing methods, Chapter 5 includes concrete applications of the methods in different case studies.

4.1 Numerical weather prediction and the need for statistical postprocessing

The goal of weather forecasting is to describe the future state of the atmosphere, which is a complex, physical system that can be modeled using a set of fundamental differential equations, referred to as the *primitive equations*. From a mathematical point of view, the

problem of weather prediction is an initial value problem, which cannot be solved analytically for the underlying system of nonlinear partial differential equations. Therefore, numerical models are applied to solve the problem, a process referred to as *numerical weather prediction* (NWP). While the underlying mathematical problem is clearly defined, the development of an NWP model includes a multitude of choices to make. In this section, we will describe the challenges of NWP that result in the need for statistical postprocessing. We refer to Warner (2010) for a detailed description of NWP and to Bauer et al. (2015) for a review on the past, present and future of NWP.

One of the biggest challenges in weather prediction is that the atmosphere is a chaotic system meaning that its predictability is inherently limited, even though an exact mathematical formulation of the system is available. Lorenz (1963), one of the founders of the *chaos theory*, describes that smallest changes in the initial values amplify over time and result in unpredictable behavior, an effect that was later famously coined as the *butterfly effect*. Next to chaos theory, there are more practical challenges to weather prediction. In order to represent the continuous meteorological variables modeled by the differential equations, NWP models operate on a discretization of time and space that depends on the model at hand. In general, NWP models become more accurate, the finer the resolution. However, the finer the resolution, the more computationally expensive the prediction becomes. The refinement of the numerical grid is therefore limited by the computational resources available. Due to the discretization, some physical processes cannot be directly resolved by the model, e.g., processes on a small scale within the grid cells, and need to be incorporated by so-called *parameterizations*. The prediction of meteorological variables becomes the more challenging, the more of the processes involved are affected by parameterizations. Wind gusts and solar irradiance, which are the main predictands in the case studies of Chapter 5, are two examples of meteorological variables that are more challenging to predict due to the parameterizations involved. In theory, the initial values of the numerical problem are given by the current state of the atmosphere at the grid points when the models are initialized. In practice, it is however not feasible to measure the desired meteorological variables at the given points in time and space. Instead, observations from weather stations, airplanes, satellites and other sources are gathered to approximate the current state of the atmosphere, a process referred to as *data assimilation*. Methods of data assimilation combine those observations with first guesses of the NWP model, while also taking into account other sources of information, in order to generate the initial values of the numerical models. Due to the limited predictability of chaotic systems and other sources of uncertainty in NWP models such as the discretization, parameterizations and the data assimilation, the necessity to quantify the uncertainty in weather prediction becomes apparent. The idea behind the aforementioned ensemble forecasts is to generate a set of equally likely future scenarios by running the NWP model multiple times based on perturbed initial values or different model specifications. We refer to an NWP model that generates ensemble forecasts as an *ensemble prediction system* (EPS).

In this thesis, we will focus on two NWP models from *Deutscher Wetterdienst* (DWD; German weather service), namely the *ICOsahedral Nonhydrostatic model* (ICON; Zängl et al., 2015) and the *COnsortium for Small-scale MOdeling model for Germany* (COSMO-DE; Baldauf et al., 2011). The (global) 40-member ICON-EPS is based on an icosahedral grid and has a horizontal resolution of 40 km.[1] In order to obtain more accurate predictions, DWD uses a finer resolution for their regional model over Europe, ICON-EU-EPS, which has a horizontal resolution of 13 km and is nested by the global ICON-EPS. From 2012 to 2018, the operational model of DWD over Germany was the 20-member COSMO-DE-EPS with a much finer horizontal resolution of 2.8 km, which allows to resolve more processes relevant for the prediction of wind gusts.[2]

Although the quality of ensemble forecasts has continuously increased since their introduction (Bauer et al., 2015), they are subject to systematic errors. The first typical type of error is a systematic bias of the ensemble members, meaning that they frequently over- or underforecast the variable of interest. An example of a biased ensemble forecast is given in the left panel of Figure 4.1. Assessing the calibration of ensemble forecasts typically reveals that the forecasts are strongly underdispersed resulting in U-shaped verification rank histograms. This can be explained by the fact that the ensemble spread is too small, that is, the forecast is too sharp. The right panel in Figure 4.1 illustrates such an dispersion error for exemplary COSMO-DE-EPS forecasts of wind gusts. In general, it can be said that ensemble forecasts do not adequately represent the forecast uncertainty. Note that we verify the ensemble forecasts against station observations in our work and that, depending on the local characteristics of the station, the degree of the systematic errors might therefore be larger.

Due to the systematic errors exhibited by ensemble forecasts, statistical postprocessing is required to achieve accurate and reliable probabilistic forecasts. Methods for statistical postprocessing are based on techniques from statistical learning and aim to correct the systematic errors of the ensemble forecasts based on a set of past predictions and observations. Figure 4.1 illustrates how biased or underdispersed ensemble forecasts can be corrected using statistical postprocessing for exemplary forecasts from the case study on wind gust prediction in Section 5.3.

[1] Here, we refer to the discretization that was used before an model update on 23 November 2022, when the grid of the ICON-EPS and ICON-EU-EPS was refined (`https://www.dwd.de/DE/leistungen/opendata/neuigkeiten/opendata_mai2022_1.html`).

[2] In 2018, the horizontal resolution of the COSMO-DE model changed to 2.2 km and the model was afterwards referred to as COSMO-D2 (Deutscher Wetterdienst, 2018). Three years later, the COSMO-D2 model was succeeded by the ICON-D2 model, which integrated the regional model for Germany in the ICON framework, also with a horizontal resolution of 2.2 km (Deutscher Wetterdienst, 2021).

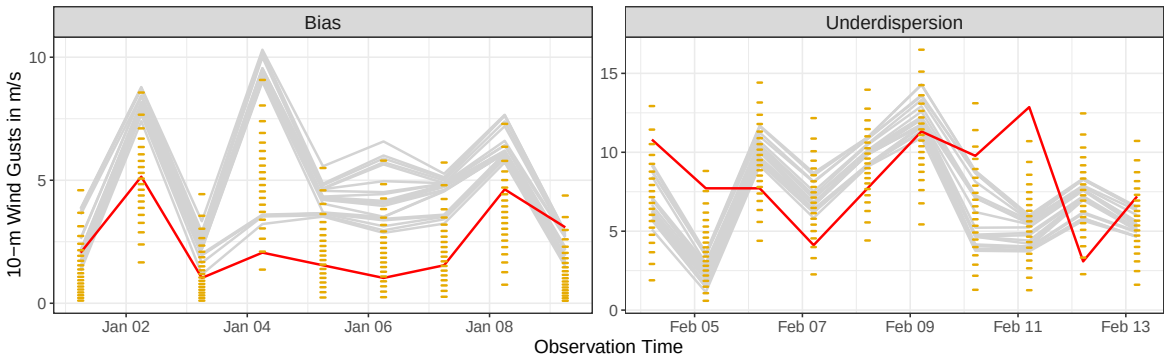

Figure 4.1: Illustration of ensemble forecasts that are subject to a bias (left) and underdispersion (right). The exemplary time series of COSMO-DE-EPS forecasts (grey), station observations (red) and statistically postprocessed forecasts (yellow) are taken from the case study on wind gust prediction in Section 5.3. The left panel is based on 6 hour forecasts for the station Mühlacker in January 2016, the right is based on 5 hour forecasts for the station Stötten in February 2016. The postprocessed forecasts were generated using EMOS, a method that will be presented in Section 4.2.1. The yellow dashes correspond to quantile forecasts at the levels $1/21, \ldots, 20/21$ derived from the predictive distribution they represent.

4.2 Established methods for postprocessing

We start the review of statistical postprocessing methods with established benchmark methods of different complexity that are in use at weather services. We use the prediction of the speed of wind gusts as running example, as it is the main application in the thesis.

We begin by introducing the notation. The weather variable of interest will be denoted by Y when we refer to the associated random variable and by y when we refer to the observed value. In our running example, we consider the speed of wind gusts which only takes positive values, therefore $y > 0$. Further, the ensemble forecasts of the variable of interest will be denoted by x. Note that $x = (x_1, \ldots, x_m)$ is an m-dimensional vector, where m is the ensemble size and x_i the ith ensemble member for $i = 1, \ldots, m$. The ensemble mean of x is denoted by $\overline{x}$ and the standard deviation by $s(x)$.

We will use the term predictor or feature interchangeably to denote a predictor variable that is used as an input to a postprocessing model, while we will refer to the vector including all predictors by $\boldsymbol{x} \in \mathbb{R}^p$, where p is the number of predictors and $\boldsymbol{x}_i$ is the ith predictor for $i = 1, \ldots, p$. For most meteorological variables, we will typically use the ensemble mean as predictor. A set of past observations, ensemble forecasts of the variable of interest and predictor vectors is denoted by $\{(y_1, x_{\cdot 1}, \boldsymbol{x}_{\cdot 1}), \ldots, (y_n, x_{\cdot n}, \boldsymbol{x}_{\cdot n})\}$, where n is the size of the set, $x_{\cdot j} = (x_{1j}, \ldots, x_{mj})$ and $\boldsymbol{x}_{\cdot j} = (\boldsymbol{x}_{1j}, \ldots, \boldsymbol{x}_{pj})$ for $j = 1, \ldots, n$.

4.2.1 Ensemble model output statistics

Ensemble model output statistics (EMOS), originally proposed by Gneiting et al. (2005) and sometimes referred to as nonhomogeneous regression, is one of the most prominent statistical postprocessing methods. EMOS is a distributional regression approach, which assumes that, given the predictor vector $\boldsymbol{x}$, the weather variable of interest Y follows a parametric distribution $\mathcal{L}(\theta)$ with $\theta \in \Theta$, where Θ denotes the parameter space of $\mathcal{L}$. The distribution parameter (vector) θ is connected to the ensemble forecast via a link function g such that

$$Y \mid \boldsymbol{x} \sim \mathcal{L}(\theta), \quad \theta = g(\boldsymbol{x}) \in \Theta. \tag{4.1}$$

Typically, the predictor variables are given by the ensemble members $x_1, \ldots, x_m$, or summary statistics of the ensemble such as the mean $\overline{x}$ and the standard deviation $s(x)$, which are linked to the parameter (vector) θ via a linear transformation depending on a set of regression coefficients that are estimated via optimum score estimation using the CRPS or LogS.

The choice of the parametric family for the forecast distribution depends on the weather variable of interest. Gneiting et al. (2005) use a Gaussian distribution for temperature and sea level pressure forecasts. More complex variables like precipitation or solar irradiance have been modeled via zero-censored distributions, whose mixed discrete-continuous nature enables point masses for the events of no rain or no irradiance.[3] In contrast to these variables, wind speed is assumed to be strictly positive and modeled via distributions that are left-truncated at zero. In the extant literature, positive distributions that have been employed include truncated normal (Thorarinsdottir and Gneiting, 2010), truncated logistic (Scheuerer and Möller, 2015), log-normal (Baran and Lerch, 2015) or truncated generalized extreme value (GEV; Baran et al., 2021b) distributions. While the differences observed in terms of the predictive performance for different distributional models are generally only minor, combinations or weighted mixtures of several parametric families have been demonstrated to improve calibration and forecast performance for extreme events (Lerch and Thorarinsdottir, 2013; Baran and Lerch, 2016, 2018). While the parametric families employed for wind speed can be assumed to be appropriate to model wind gusts as well, specific studies tailored to wind gusts are scarce. For our main application on wind gust predictions, we therefore focus on the distributions applied to wind speed. In agreement with Pantillon et al. (2018), preliminary results for the comparison of the distribution types used for wind speed for data from the case study on wind gust forecasting in Section 5.3 have shown that the differences among them are minor. In the following, we will focus on the truncated logistic distribution, which was introduced in Section 2.3.1.

After identifying a distribution for our EMOS model, the location parameter $\mu \in \mathbb{R}$ and scale parameter $\sigma > 0$ need to be linked to the ensemble forecast x. We are using the link

[3] In the case studies in Sections 5.1 and 5.2, we will model precipitation and solar irradiance variables with the zero-censored logistic distribution, which was introduced in Section 2.3.1.

function

$$g(x; a, b, c, d) := (\mu(x; a, b), \sigma(x; c, d)), \tag{4.2}$$

where

$$\mu(x; a, b) := a + \exp(b)\,\overline{x}, \quad \sigma(x; c, d) := \exp(c + d \log(s(x))) \tag{4.3}$$

and $a, b, c, d \in \mathbb{R}$ are the EMOS parameters that are estimated via optimum score estimation. At this point, we want to comment on the implementation of the EMOS approach used in the case study on wind gust predictions in Section 5.3. We start with the general distributional regression setting based on optimum score estimation, where we specify neither a forecast distribution nor a link function. Let S be the corresponding strictly proper scoring rule, then the general optimization problem we have to solve is given by

$$\min_{g \in \mathcal{H}} \bar{S}_n(g) \quad \text{with} \quad \bar{S}_n(g) := \frac{1}{n} \sum_{j=1}^{n} S(\mathcal{L}(g(\boldsymbol{x}_{\cdot j})), y_j), \tag{4.4}$$

where $\mathcal{H}$ is a suitable function space, $\mathcal{L}(\theta)$ the forecast distribution dependent on the parameter vector $\theta = g(\boldsymbol{x})$ and $\{(\boldsymbol{x}_{\cdot 1}, y_1), \ldots, (\boldsymbol{x}_{\cdot n}, y_n)\}$ a training set of size n. As aforementioned, the link function g is typically a linear function or a transformation thereof based on the ensemble members or ensemble summary statistics such as the link function used in equations (4.2) and (4.3). Denoting the coefficients of the linear combination within the link function g by $\vartheta \in \mathbb{R}^d$, where d denotes the number of coefficients, we can rewrite the problem in equation (4.4) as follows

$$\min_{\vartheta \in \mathbb{R}^d} \bar{S}_n(\vartheta) \quad \text{with} \quad \bar{S}_n(\vartheta) := \frac{1}{n} \sum_{j=1}^{n} S(\mathcal{L}(g(\boldsymbol{x}_{\cdot j}; \vartheta)), y_j). \tag{4.5}$$

Now, we return to the case of a logistic distribution left-truncated at zero and the parameterization of location and scale parameter as in equations (4.2) and (4.3). In practice, the optimization problem in equation (4.5) must be solved numerically for the CRPS and LogS, whose variants for the truncated logistic distribution have been introduced in Section 2.3.1. Typically, this is done using gradient-based optimizers. Although an analytic formula of the gradient is not required for such optimization routines, providing them reduces the computational efforts and results in closer approximations to the optimal values. Given the parameterization in equations (4.2) and (4.3), we can calculate the partial derivatives with respect to the EMOS parameter vector $\vartheta := (a, b, c, d)$:

$$\partial_a \bar{S}_n(\vartheta) = \frac{1}{n} \sum_{j=1}^{n} \partial_\mu S(\mathcal{L}(g(x_{\cdot j}; \vartheta)), y_j), \tag{4.6}$$

$$\partial_b \bar{S}_n(\vartheta) = \frac{1}{n}\sum_{j=1}^{n} \overline{x_{\cdot j}} \cdot \exp(b) \cdot \partial_\mu S\left(\mathcal{L}\left(g\left(x_{\cdot j};\vartheta\right)\right), y_j\right), \tag{4.7}$$

$$\partial_c \bar{S}_n(\vartheta) = \frac{1}{n}\sum_{j=1}^{n} \sigma\left(x_{\cdot j}; c, d\right) \cdot \partial_\sigma S\left(\mathcal{L}\left(g\left(x_{\cdot j};\vartheta\right)\right), y_j\right), \tag{4.8}$$

$$\partial_d \bar{S}_n(\vartheta) = \frac{1}{n}\sum_{j=1}^{n} \left(\log s(x_{\cdot j})\right) \cdot \sigma\left(x_{\cdot j}; c, d\right) \cdot \partial_\sigma S\left(\mathcal{L}\left(g\left(x_{\cdot j};\vartheta\right)\right), y_j\right), \tag{4.9}$$

where $\{(x_{\cdot 1}, y_1), \ldots, (x_{\cdot n}, y_n)\}$ is a training set of size n. The gradient of both the CRPS and LogS can be calculated analytically for truncated or censored logistic distributions (Jordan et al., 2019). Hence, we can implement a gradient-based optimization in order to solve the optimization problem underlying our EMOS variant based on the partial derivatives in equations (4.6)–(4.9).

4.2.2 Gradient-boosting extension of EMOS

The general distribution regression framework in equation (4.1) allows for any real-valued predictor vector $\boldsymbol{x}$, still most EMOS approaches rely only on forecasts of the weather variable of interest or a small selection of carefully chosen variables, although a large variety of additional meteorological predictor variables is available. For the fact that the classic EMOS approach from Section 4.2.1 is not appropriate for a (large) set of multiple predictor variables, there are two main explanations, which mirror problems from deterministic regression settings such as the linear least squares regression.

The first is the specification of the underlying link function g, which connects the predictor variables and the distribution parameters. Based on the assumption that ensemble forecasts are biased and subject to dispersion errors such as underdispersion, a linear transformation of the ensemble mean or members and of the standard deviation of the ensemble forecasts is a straightforward choice for the type of the link function. However, if we include forecasts of other meteorological variables, we have to make an assumption on the influence of this variable on the distributional parameters, which define the forecast distribution. Given the example of wind gusts, it is not straightforward to see how forecasts of temperature or precipitation are connected to the target variable. Often, nonlinear relations and interactions between different meteorological variables are present, which are difficult to incorporate in the link function. Even though a linear relation between the predictor variable and the predictand might not be supported from a meteorological point of view, using a linear link function based on a set of predictor variables might still result in superior predictive performance.

Introducing a large set of predictor variables typically results in a more complex model and one is confronted with the problem of *overfitting*, which is strongly connected to the *bias-variance trade-off* (e.g., Hastie et al., 2009). A complex model with a large degree of freedom has the capability to learn complex relations from the data, but may also (over-)adapt

to spurious patterns in the training data that are the result of data sampling, therefore the model has a large estimation variance. On the other side, a parsimonious model is not capable of adapting to complex patterns in the data, but its estimation is much more robust due to the lower number of parameters to be estimated. The first case, where a model overadapts to the training data and does not generalize well on unseen data, is referred to as overfitting. Returning back to EMOS, we face the problem of overfitting when using a large set of predictors, as the number of coefficients becomes large. If one does not want to carefully select the predictor variables by hand with the risk of finding a suboptimal set of variables, one has to make use of automated regularization techniques, i.e., modifications of the model estimation that prevent the model from overfitting. There exists a large variety of regularization techniques, especially in modern ML where complex models with enormous numbers of parameters are used, much focus lies on regularization (e.g., Hastie et al., 2009).

Here, we rely on a gradient-boosting technique for statistical postprocessing of ensemble forecasts via EMOS. Messner et al. (2017) introduce the *gradient-boosting extension of EMOS* (EMOS-GB) for distributions that are defined by a real-valued location parameter μ and a positive scale parameter $\sigma > 0$, that is, $\theta = (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_{>0}$. Examples are the normal and logistic distribution or truncated and censored variants thereof. Adaptation towards other parametric distributions is straightforward. In line with most EMOS approaches and our parameterization in equations (4.2) and (4.3), Messner et al. (2017) introduce their approach for linear parameterizations of the form

$$\mu\left(\boldsymbol{x}; a, b_1, \ldots, b_p\right) := a + \sum_{i=1}^{p} b_i \boldsymbol{x}_i, \quad a, b_1, \ldots, b_p \in \mathbb{R}, \tag{4.10}$$

$$\sigma\left(\boldsymbol{x}; c, d_1, \ldots, d_p\right) := \exp\left(c + \sum_{i=1}^{p} d_i \boldsymbol{x}_i\right), \quad c, d_1, \ldots, d_p \in \mathbb{R}. \tag{4.11}$$

Again, EMOS-GB can be easily adapted towards other (differentiable) parameterizations. EMOS-GB solves the optimum score minimization problem in equation (4.5) iteratively using a gradient-boosting algorithm. The idea of boosting is to initialize all coefficient values at zero, and to update only the coefficients of the predictor that improves the predictive performance most. Based on the gradient of the loss function that can be calculated analogously to equations (4.6)–(4.9), the predictor variables with the highest correlation to the gradients of the location and scale parameter are selected. Then, that of the two updates that improves the current fit most is applied, where update refers to taking a step in the direction of the steepest descent. The size of the step is fixed and predetermined by the user as a hyperparameter. This procedure is repeated until a stopping criterion such as a maximum number of iterations or a threshold for the improvement of the fit is reached.

4.2.3 Member-by-member postprocessing

Member-by-member postprocessing (MBM; van Schaeybroeck and Vannitsem, 2015) is based on the idea to adjust each member individually in order to generate a calibrated ensemble. Wilks (2018) highlights several variants of MBM, which are of the general form

$$\tilde{x}_i := (a + b\overline{x}) + \gamma\,(x_i - \overline{x})\,, \quad i = 1, \ldots, m, \quad a, b, \gamma \in \mathbb{R}, \tag{4.12}$$

where $\tilde{x} = (\tilde{x}_1, \ldots, \tilde{x}_m)$ denotes the postprocessed ensemble. The first term in equation (4.12) represents a bias-corrected ensemble mean, whereas the second term includes the individual members with γ scaling the distance to the ensemble mean.

Note that the mean of the adjusted ensemble depends only on the parameters a and b and not the stretch coefficient γ:

$$\begin{aligned} \overline{\tilde{x}} &= \frac{1}{m}\sum_{i=1}^{m}\left[(a + b\overline{x}) + \gamma\,(x_i - \overline{x})\right] \\ &= (a + b\overline{x}) + \gamma\left(\frac{1}{m}\sum_{i=1}^{m} x_i - \overline{x}\right) \\ &= a + b\overline{x} \end{aligned} \tag{4.13}$$

Our implementation of MBM postprocessing follows van Schaeybroeck and Vannitsem (2015), who let the stretch coefficient γ depend on the *ensemble mean difference* via

$$\gamma := c + \frac{d}{\delta(x)}, \tag{4.14}$$

where $c, d \in \mathbb{R}$ and the ensemble mean difference δ is defined by

$$\delta(x) := \frac{1}{m^2}\sum_{i,l=1}^{m} |x_i - x_l|\,. \tag{4.15}$$

Note that the definition of the stretch coefficient in equation (4.14) results in an ensemble mean difference of the adjusted ensemble that is a linear transformation of that of the original ensemble, as the following calculation shows:

$$\begin{aligned} \delta(\tilde{x}) &= \frac{1}{m^2}\sum_{i,l=1}^{m} |\tilde{x}_i - \tilde{x}_l| \\ &= \frac{1}{m^2}\sum_{i,l=1}^{m} |\gamma\,(x_i - x_l)| \\ &= |\gamma|\,\delta(x) \\ &= |c\delta(x) + d|\,. \end{aligned} \tag{4.16}$$

Hence, the parameters c and d can be used to adjust the ensemble mean difference of the MBM ensemble forecast, which is a measure of the sharpness of the forecast distribution.

Following the paradigm of Gneiting et al. (2007), we want to estimate the MBM parameters via optimum score estimation. However, the MBM forecasts are given in form of an ensemble and not a full predictive distribution. Therefore, one either has to resort on assumptions for the distribution of the ensemble forecasts or estimate the parameters in a distribution-free approach. Here, we consider two techniques presented in Wilks (2018), which include one approach based on MLE and one on CRPS estimation.

The first approach is based on the assumption that the AE of the mean of the adjusted ensemble follows an exponential distribution with mean $\delta(\tilde{x})$, that is, the ensemble mean difference of the adjusted ensemble. Altogether, that is,

$$|\overline{\tilde{x}} - Y| \mid x \sim \operatorname{Exp}\left(\delta\left(\tilde{x}\right)^{-1}\right). \tag{4.17}$$

Under this distributional assumption, we can now calculate the LogS for a given observation $y \in \mathbb{R}$. Let f be the PDF of an exponential distribution with mean $\delta(\tilde{x})$, then

$$\begin{aligned}
\operatorname{LogS}\left(f, |\overline{\tilde{x}} - y|\right) &= -\log\left(f(|\overline{\tilde{x}} - y|)\right) \\
&= -\log\left(\delta\left(\tilde{x}\right)^{-1} \exp\left(-\frac{|\overline{\tilde{x}} - y|}{\delta\left(\tilde{x}\right)}\right)\right) \\
&= \frac{|\overline{\tilde{x}} - y|}{\delta\left(\tilde{x}\right)} + \log \delta\left(\tilde{x}\right) \\
&= \frac{|(a + b\overline{x}) - y|}{|c\delta(x) + d|} + \log|c\delta(x) + d|\,.
\end{aligned} \tag{4.18}$$

Note that the LogS does not depend on the accuracy of the adjusted ensemble members but only on the accuracy of the corresponding mean forecast. Analogous to the optimization problem in equation (4.5), the MBM parameters are estimated by minimizing the mean LogS over the training set.

The other variant for the estimation of the MBM parameters is a distribution-free approach based on the sample CRPS. The advantage of this method is that by using the empirical distribution function to calculate the CRPS no assumption on the distribution is required. Using the expectation representation of the CRPS in equation (2.11), the CRPS for the empirical distribution function $\hat{F}$ based on the sample $\tilde{x} = (\tilde{x}_1, \ldots, \tilde{x}_m)$ is given by

$$\operatorname{CRPS}\left(\hat{F}, y\right) = \frac{1}{m}\sum_{i=1}^{m} |\tilde{x}_i - y| - \frac{1}{2}\delta(\tilde{x}), \quad y \in \mathbb{R}, \tag{4.19}$$

according to Jordan et al. (2019). The accuracy of each ensemble member is penalized in the first term, the second term penalizes the sharpness via the ensemble mean difference. In line with the MLE approach, the MBM parameters can be estimated by minimizing the

mean CRPS on the training set. Note that the calibrated ensemble $\tilde{x}$ depends on the MBM parameters via equations (4.12) and (4.14).

A main advantage of MBM compared to all other approaches is that the rank correlation structure of the ensemble forecasts is preserved by postprocessing, since each member is transformed individually by the same linear transformation. MBM thus results in forecasts that are physically consistent over time, space and also different weather variables, even if MBM is applied for each component separately (van Schaeybroeck and Vannitsem, 2015; Schefzik, 2017; Wilks, 2018).

4.2.4 Isotonic distributional regression

Henzi et al. (2021b) propose *isotonic distributional regression* (IDR), a novel nonparametric regression technique, which results in simple and flexible probabilistic forecasts as it depends neither on distributional assumptions nor prespecified transformations. Since it requires no parameter tuning and minimal implementation choices, it is an ideal generic benchmark in probabilistic forecasting tasks.

IDR is built on the assumption of an isotonic relationship between the predictors and the target variable. Given a partial order $\preceq$ on the covariate space, IDR generates probabilistic forecasts such that $F(\boldsymbol{x}) \preceq_{st} F(\boldsymbol{x'})$ if $\boldsymbol{x} \preceq \boldsymbol{x'}$, where $F(\boldsymbol{x})$ and $F(\boldsymbol{x'})$ denote the IDR forecasts dependent on the predictor vectors $\boldsymbol{x}$ and $\boldsymbol{x'}$ respectively and $\preceq_{st}$ is the *stochastic order* defined by

$$F \preceq_{st} G \qquad :\Longleftrightarrow \qquad F(y) \geq G(y) \quad \forall y \in \mathbb{R}, \tag{4.20}$$

for two probability measures F and G. In the univariate case with only one predictor, isotonicity is based on the linear ordering on the real line, that is, $z \preceq z'$ if, and only if, $z \leq z'$ for $z, z' \in \mathbb{R}$. When multiple predictors (such as multiple ensemble members) are given, the multivariate covariate space is equipped with a partial order. Then, the only implementation choice required for IDR is the selection of a partial order on the covariate space. Among the choices introduced in Henzi et al. (2021b), the *componentwise order* defined by

$$z \preceq_{comp} z' \qquad :\Longleftrightarrow \qquad z_i \leq z'_i, \quad i = 1, \ldots, p, \tag{4.21}$$

for $z \in \mathbb{R}^p$ is not appropriate under the assumption of exchangeability among the ensemble members. On the other side, the *empirical stochastic order*

$$z \preceq_{sd} z' \qquad :\Longleftrightarrow \qquad \hat{F}(z) \preceq_{st} \hat{F}(z'), \quad z \in \mathbb{R}^p, \tag{4.22}$$

where $\hat{F}(z)$ denotes the empirical distribution of $z \in \mathbb{R}^p$, and the *empirical increasing convex*

order, which we define for $z \in \mathbb{R}^p$ via

$$z \preceq_{icx} z' \qquad :\Longleftrightarrow \qquad \sum_{i=l}^{p} z_i \leq \sum_{i=l}^{p} z'_i, \quad l = 1, \ldots, p, \tag{4.23}$$

are appropriate for the situation at hand when all ensemble members are used as predictors for the IDR model. Note that the empirical stochastic order is equivalent to the componentwise order on the order statistics of the predictor vector and is a stronger notion than the empirical increasing convex order, meaning that the stronger order implies the weaker.

Under those order restrictions, a conditional distribution that is optimal with respect to a broad class of relevant loss functions including proper scoring rules is then estimated. Conceptually, IDR can be seen as a far-reaching generalization of widely used isotonic regression techniques that are based on the PAV-algorithm (de Leeuw et al., 2009). To the best of our knowledge, the case study on probabilistic wind gust forecasting in Section 5.3 is the first application of IDR in a postprocessing context besides the case study on precipitation accumulation in Henzi et al. (2021b), who find that IDR forecasts were competitive to EMOS.

4.2.5 Quantile regression forests

A nonparametric, data-driven technique that neither relies on distributional assumptions, link functions nor parameter estimation is *quantile regression forests* (QRF), which was first used in the context of postprocessing by Taillardat et al. (2016). RFs are randomized ensembles of decision trees, which operate by splitting the predictor space in order to create an analog forecast (Breiman, 1984). This is done iteratively by first finding an order criterion based on the predictor that explains the variability within the training set best, and then splitting the predictor space according to this criterion. This procedure is repeated on the resulting subsets until a stopping criterion is reached, thereby creating a partition of the predictor space. Following the decisions at the so-called nodes, one then obtains an analog forecast based on the training samples. RFs create an ensemble of decision trees by considering only a randomly chosen subset of the training data at each tree and of the predictors at each node, aiming to reduce correlation between individual decision trees (Breiman, 2001). QRF extends the RF framework by performing a quantile regression that generates a probabilistic forecast (Meinshausen, 2006). The QRF forecast thus approximates the forecast distribution by a set of quantiles derived from the set of analog observations.

4.3 Neural network-based postprocessing

Over the past decade, NNs have become ubiquitous in data-driven scientific disciplines and have in recent years been increasingly used in the postprocessing literature (see Vannitsem et al., 2021, for a recent review). NNs are universal function approximators for which a variety

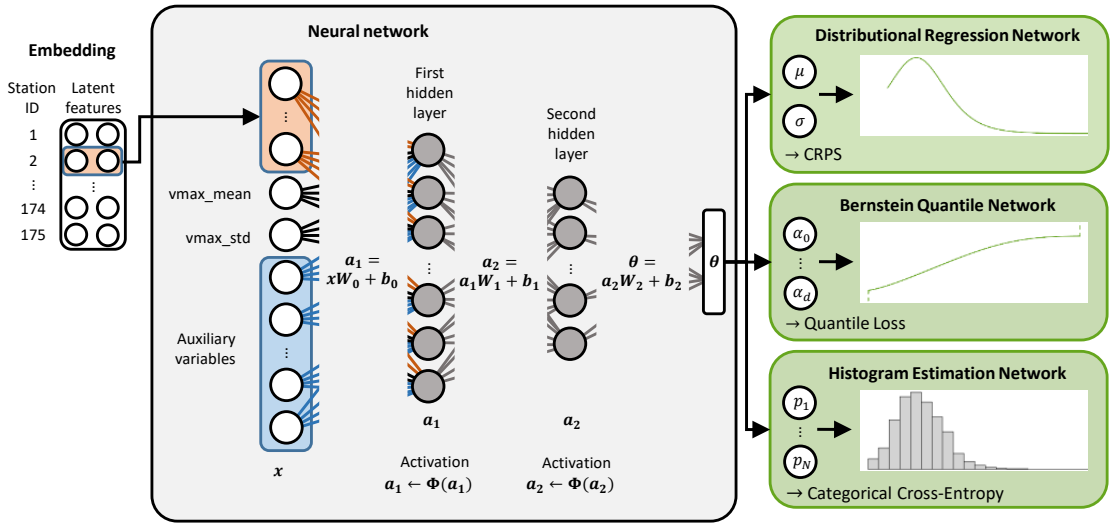

Figure 4.2: Graphical illustration of the framework for NN-based postprocessing presented in Section 4.3.1.

of highly complex extensions has been proposed. However, NN models often require large datasets and computational efforts, and are sometimes perceived to lack interpretability.

In the following, we will present a group of postprocessing methods based on NNs that has already been investigated regarding the aggregation of distributional forecasts from deep ensembles in Chapter 3, but not been properly introduced yet. Following the introduction of a general framework of our network-based postprocessing methods, we will introduce three model variants that are based on the forecast distributions introduced in Section 2.3. In the interest of brevity, we will assume a basic familiarity with NNs and the underlying terminology. We refer to McGovern et al. (2019) for an accessible introduction in a meteorological context and to Goodfellow et al. (2016) for a detailed review.

4.3.1 A framework for neural network-based postprocessing

The use of NNs in a distributional regression-based postprocessing context was first proposed in Rasp and Lerch (2018). Our framework for NN-based postprocessing builds on their approach and subsequent developments in Bremnes (2020), Scheuerer et al. (2020), and Veldkamp et al. (2021), among others. In particular, we propose three model variants that utilize a common basic NN architecture, but differ in terms of the form that the probabilistic forecasts take which governs both the output of the NN as well as the loss function used for parameter estimation. A graphical illustration of this framework is presented in Figure 4.2.

The rise of artificial intelligence and NNs is closely connected to the increase in data availability and computing power, as these methods unfold their strengths when modeling complex nonlinear relations trained on large datasets. A main challenge in the case of

postprocessing is to find a way to optimally utilize the entirety of available input data while preserving the inherent spatial and temporal information. We focus on building one network jointly for all stations at a given lead time, which we will refer to as locally adaptive joint network. For this purpose, Rasp and Lerch (2018) propose a station embedding, where a station identifier is mapped to a vector of latent features, which are then used as auxiliary input variables of the NN. The estimation of the embedding mapping is integrated into the overall training procedure and aims to model local characteristics implicitly, contrary to Lerch and Baran (2017) and Hamill et al. (2008), who apply a preliminary procedure to pool stations respectively grid points that exhibit similar characteristics.

Our basic NN architecture consists of 2 hidden layers and a customized output. The training procedure is based on the *adaptive moment estimation algorithm* (Adam; Kingma and Ba, 2014), and the weights of the network are estimated on the training period by optimizing a suitable loss function tailored to the desired output. We apply an early-stopping strategy that stops the training process when the validation loss remains constant for a given number of epochs to prevent the model from overfitting.

Chapter 3 is motivated by the fact that NN-based forecasting models are often run several times from randomly initialized weights and batches to produce deep ensembles in order to account for the randomness of the training process based on stochastic gradient descent methods. We follow this principle and produce an ensemble of 10 models for each of the three variants of NN-based postprocessing models introduced above, which leads to the question how the resulting deep ensembles should be aggregated into a single probabilistic forecast for every model. Following the conclusions in Chapter 3, we combine the predictive distributions using VI, that is, quantile aggregation.[4] We refer to Section 3.2 for a detailed description of the aggregation of the forecast distributions associated with the three network variants.

4.3.2 Distributional regression network

Rasp and Lerch (2018) propose a postprocessing method based on NNs that extends the EMOS framework in equation (4.1), which we coined DRN in Section 3.2.1. A key component of the improvement is that instead of relying on prespecified link functions such as equations (4.2) and (4.3) or (4.10) and (4.11) to connect input predictors to distribution parameters, an NN is used to model a flexible and nonlinear relation in a data-driven way. Given a specific network architecture that defines a total of d weights (and biases) $\vartheta \in \mathbb{R}^d$, the NN can be interpreted as a link function $g(\cdot; \vartheta)$. Then, the estimation of the NN corresponds to an optimization problem that can be written analogously to equation (4.5), where S is a strictly

[4] The question of finding the best aggregation method for the deep ensembles presented in the case study in Section 5.3 was the starting point for Chapter 3. While developing the models for the case study in Section 5.3, we found that VI was preferable over LP for all three variants. However, at that point, we did not investigate the optimal size of the deep ensemble. Instead, we followed Rasp and Lerch (2018) and Bremnes (2020) who use deep ensembles of size 10. Only later, we conducted the case study in Section 3.4, which focuses on the aspect of aggregation, and found that 10 was also the optimal ensemble size.

proper scoring rule such as the CRPS or LogS. However, the problem is solved differently than in the EMOS approach using stochastic gradient descent methods as is standard practice for NNs. The output of the NN is thus given by the forecast distribution parameters $\theta = g(\boldsymbol{x})$ for a predictor vector $\boldsymbol{x}$.

4.3.3 Bernstein quantile network

Bremnes (2020) extends the DRN framework of Rasp and Lerch (2018) towards a semiparametric approach based on the Bernstein quantile function introduced in Section 2.3.2 such that

$$Y \mid \boldsymbol{x} \sim Q_\alpha, \quad \alpha = g(\boldsymbol{x}) \in \mathbb{R}^{d+1}, \tag{4.24}$$

where the link function g is given by the NN and Q_α denotes the Bernstein quantile function in equation (2.33) dependent on the coefficient vector $\alpha = (\alpha_0, \ldots, \alpha_d)$. The Bernstein quantile network (BQN; Section 3.2.2) was proposed for wind speed forecasting, but can be readily applied to other target variables of interest due to its flexibility. The key implementation choice is the degree d, with larger values leading to a more flexible forecast but also a larger estimation variance. For a given degree d, the BQN forecast is fully defined by the $d+1$ basis coefficients. In contrast to Bremnes (2020), we enforce monotonicity by targeting the increments $\tilde{\alpha}_0 \in \mathbb{R}$ and $\tilde{\alpha}_l \geq 0$, $l = 1, \ldots, d$, based on which the coefficients can be derived via the recursive formula

$$\alpha_0 = \tilde{\alpha}_0, \quad \alpha_l = \alpha_{l-1} + \tilde{\alpha}_l, \quad l = 1, \ldots, d. \tag{4.25}$$

We obtain the increments as output of the NN and apply a softplus activation function in the output layer (besides the first component), which ensures positivity of the increments and, according to Section 2.3.2, thereby strictly increasing coefficients and quantile functions. As noted in Section 2.3.2, the lower bound of the support is given by α_0, hence the softplus-activation function should only be applied to the first component of the output when we assume that the target variable is positive. Due to the lack of a readily available closed form expression of the CRPS or LogS for BQN forecasts and following Bremnes (2020), the network parameters are estimated based on the QS. According to equation (2.20), we can approximate the CRPS with the average QS over a sufficiently large set of equidistant quantile levels. Hence, we use the following loss function for our network:

$$S(Q_\alpha, y; n_q) := \frac{1}{n_q} \sum_{i=1}^{n_q} \rho_{\tau_i}(Q_\alpha(\tau_i), y) \quad \text{with} \quad \tau_i = \frac{i}{n_q + 1} \quad \text{for} \quad i = 1, \ldots, n_q, \tag{4.26}$$

where Q_α denotes the BQN forecast, $y \in \mathbb{R}$ the observation and n_q the number of equidistant quantiles considered.

4.3.4 HISTOGRAM ESTIMATION NETWORK

The third network-based postprocessing method may be considered as a universally applicable approach to probabilistic forecasting and is based on the idea to transform the probabilistic forecasting problem into a classification task, one of the main applications of NNs. This is done by partitioning the observation range in distinct classes and assigning a probability to each of them. In mathematical terms, this is equivalent to assuming that the probabilistic forecast is given by the piecewise uniform distribution introduced in Section 2.3.3, that is,

$$Y \mid \boldsymbol{x} \sim \mathcal{U}(b, p), \quad (b, p) = g(\boldsymbol{x}) \in \mathbb{R}^{2N+1}, \tag{4.27}$$

where $\mathcal{U}(b, p)$ denotes a piecewise uniform distribution with N bins based on the edges $b = (b_0, \ldots, b_N)$ and bin probabilities $p = (p_1, \ldots, p_N)$. Because the PDF of such a distribution is given by a piecewise constant function, which resembles a histogram, we refer to this approach as histogram estimation network (HEN; Section 3.2.3). Variants of this approach have been used in a variety of disciplines and applications (e.g., Felder et al., 2018; Gasthaus et al., 2019; Li et al., 2021). For recent examples in the context of postprocessing, see Scheuerer et al. (2020) and Veldkamp et al. (2021).

Given a fixed number of bins specified as a hyperparameter, the bin edges and corresponding bin probabilities need to be determined. There exist several options for the output of the NN architecture to achieve this. The most flexible approach, for example implemented in Gasthaus et al. (2019), would be to obtain both the bin edges and the probabilities as output of the NN. We here instead follow a more parsimonious alternative and fix the bin edges, so that only the bin probabilities are determined by the NN, which can be interpreted as a probabilistic classification task. Note that alternatively, it is also possible to fix the bin probabilities and determine the bin edges by the NN. This would be equivalent to estimating the quantiles of the forecast distribution at the levels defined by the prespecified probabilities, that is, a quantile regression.

Both the CRPS and the LogS, which we have calculated in Section 2.3.3, can be used to estimate the NN. As stated in Section 2.3.3, minimizing the LogS reduces to the categorical cross-entropy in equation (2.57) for our approach with fixed bins. Thus, one can estimate the HEN forecast via a standard classification network preserving the optimum scoring principle.

CHAPTER 5

Statistical Postprocessing: Case Studies

Facing the decision to select a method for statistical postprocessing, the most sophisticated model may not always be the best choice. The more complex a method is or the more degrees of freedom it includes, the more data is required, in a sense that not only the sample size but also the inherent information increases, a well-known principle in statistics (e.g., Hastie et al., 2009, Chapter 7). Hence, practical aspects ought to be taken into account and a model should be tailored to the situation at hand in order to result in the best predictive performance.

The methods presented in Chapter 4 can be divided in three groups of increasing complexity. Starting with established, comparatively simple techniques rooted in statistics, EMOS, MBM and IDR form the first group of basic statistical postprocessing techniques, where the term basic refers to the fact that these methods solely use the ensemble forecasts of the variable of interest as predictors. The second group of postprocessing methods consists of the benchmark ML methods QRF and EMOS-GB that are able to incorporate additional predictor variables besides ensemble forecasts of the variable of interest in an automated, data-driven way. The third group of postprocessing methods comprises the NN variants DRN, BQN and HEN.

Following the introduction of the methods in Chapter 4, we are applying them in three case studies of increasing complexity here in Chapter 5. The first introductory section exemplifies the effects of statistical postprocessing using the basic EMOS approach. Building up on that, the second section includes more sophisticated EMOS models for solar irradiance, which is a nonstandard forecasting target and has in recent times received more attention due to its importance for renewable energy sources. The third and last case study is the most comprehensive, a systematic comparison of the three groups for wind gust forecasting. In each case study, we will introduce the setting, present the data, describe the model configurations tailored to the situation at hand and then assess the predictive performance guided by the principle of maximizing sharpness subject to calibration. At last, we wrap up the chapter discussing advantages and shortcomings of the postprocessing methods based on the conclusions drawn from the case studies.

5.1 NEAR REAL-TIME POSTPROCESSING ON KIT-WEATHER

The Department Troposphere Research of the Institute of Meteorology and Climate Research (IMK-TRO) at the Karlsruhe Institute of Technology (KIT) runs the *KIT-Weather* portal, where several forecasting products have been implemented by members of IMK-TRO.[1] One of the products are the so-called *ensemble boxplot meteograms*, which illustrate the temporal evolution of operational ICON ensemble forecasts in near real-time via boxplots. The user can choose between meteograms of several meteorological variables for various European cities including lead times up to six and a half days.

To test statistical postprocessing methods in a pseudo-operational setting, we have implemented a near real-time postprocessing approach on the portal. We decided to use the robust EMOS approach presented in Section 4.2.1, since it can be easily adapted towards several of the available meteorological variables and handles small amounts of data well due to its parsimony. As the postprocessed forecasts, which are accessible on the portal, can be compared directly to the corresponding ensemble forecasts, they illustrate the effects of postprocessing on different types of ensemble forecasts. Hence, they also serve one of the main purposes of the portal, that is, dissemination and outreach. The portal operates in a near real-time setting meaning that the ensemble forecasts from the latest model initialization are automatically downloaded, displayed and postprocessed on the server.

Figure 5.1 displays a screenshot of the portal including meteograms of the raw model output and the corresponding postprocessed forecasts. Note that the boxplots also illustrate the quantiles at the 10% and 90% level, and that the whiskers extend to the minimum and maximum of the ensemble.[2] While the boxplots are based on a 40-member ensemble, EMOS generates a full predictive distribution. For the illustration of the postprocessed forecasts, we replace the ensemble with 40 equidistant predictive quantile forecasts at levels $1/41, \ldots, 40/41$. Comparing the meteogram before and after postprocessing, the most obvious finding is that the length of the boxes increases resulting in less sharp forecasts, a behavior that will be explained in Section 5.1.3, where we evaluate the postprocessed forecasts.

5.1.1 DATA

At the beginning of the project, the portal included eight meteorological variables, namely, temperature, pressure, wind speed, precipitation rate and sum, cloud cover as well as direct and diffuse solar irradiance.[3] As mentioned at the beginning of the chapter, solar irradiance

[1] `http://www.kit-weather.de/`

[2] For all other boxplots in this thesis, the whiskers reach out at most 1.5 times the interquartile range, as is standard.

[3] Additional variables such as wind gusts, snow rate or integrated water vapour were integrated at a later point in time. Of these variables, postprocessing was later implemented only for wind gusts as the other variables do not fit the postprocessing framework presented in this section and more data is required. Wind gusts are however not considered in this section, since the underlying dataset is substantially smaller than

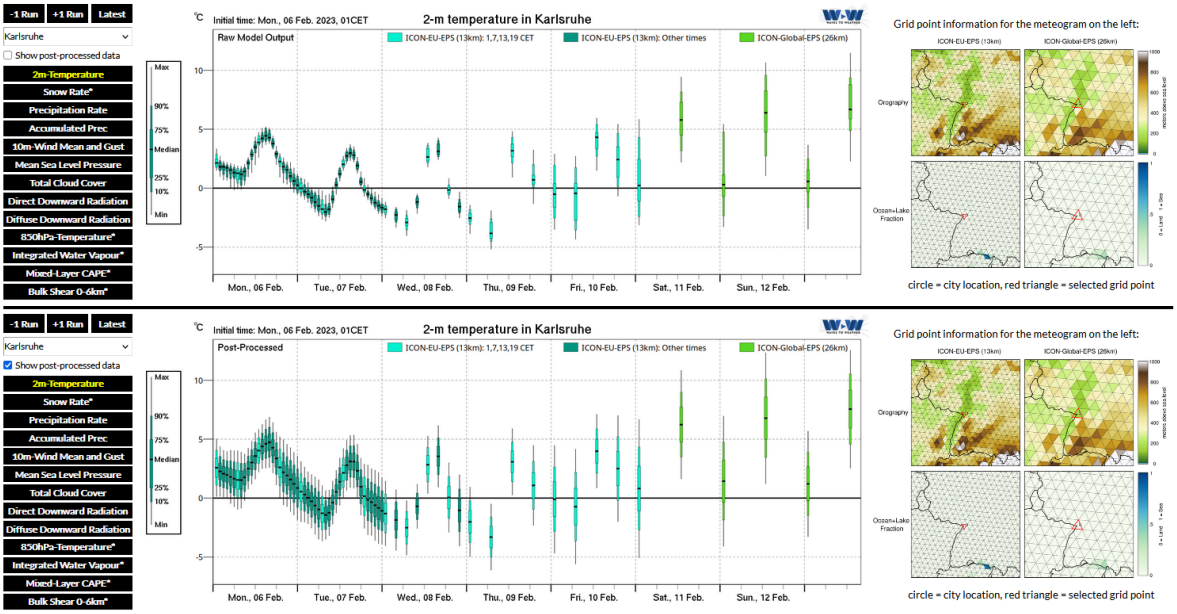

Figure 5.1: Screenshot of the ensemble boxplot meteograms on the KIT-Weather portal for temperature in Karlsruhe (taken on 6 February 2023). The ensemble (top) and postprocessed forecasts (bottom) are based on the ICON model run initialized at 00 UTC (01 CET) on 6 February 2023. On the left, the model run, the location, the variable and postprocessing can be selected. On the right, information on the grid point of the corresponding location is available.

variables are nonstandard for statistical postprocessing, we thus cover those two variables in more detail in Section 5.2. The variables we consider are described in Table 5.1.

The ensemble forecasts on the portal stem from the operational ICON model that was introduced in Section 4.1. The 40-member ICON-EPS forecasts shown in the meteograms include up to 68 lead times reaching up to 180 hours. The lead times are not equidistant, instead the temporal resolution is finer for shorter lead times and becomes coarser as the lead time increases. Within the first 48 hours, the temporal resolution is 1-hourly, it is 3-hourly up to 72 hours, 6-hourly up to 120 hours and 12-hourly between 120 and 180 hours. The forecasts up to 120 hours are from the ICON-EU-EPS, which is initialized four times a day at 00, 06, 12 and 18 UTC, while the remaining forecasts are taken from the ICON-Global model, which is initialized only twice a day at 00 and 12 UTC. Note that the ICON-EPS is generated with the help of random perturbations and the ensemble members can thus be considered exchangeable. For each city, the forecasts are taken from a representative grid point of the surrounding grid cell, which is illustrated on the portal (Figure 5.1). The ICON ensemble data is downloaded from the Open Data Server of DWD and archived several times a day by IMK-TRO.[4] Archiving is required because only the most recent model runs are available.

for the other variables and the variable will be covered in Section 5.3 and Chapter 6 in detail.

[4] `https://opendata.dwd.de/`

Table 5.1: Overview of the meteorological variables on the KIT-Weather portal considered in this section, including the availability of observations at surface stations and the initial date of the forecast archive, each for the ICON-EU and ICON-Global model.

Variable	Description	Unit	KA	BE	HH	MZ	MU	Start EU	Start Global
T_2M	Temperature at 2 m	°C	✓	✓	✓	✓	✓	12/12/2018	08/01/2019
MSLP	Air pressure (mean sea level)	hPa	✓	✓	✓	-	✓	18/12/2018	-
WIND_10M	Momentary wind speed at 10 m	km/h	✓	✓	✓	-	✓	17/12/2018	08/01/2019
PREC_RATE	Total precipitation rate	mm/h	✓	✓	✓	✓	✓	17/01/2019	12/02/2019
PREC_SUM	Total precipitation sum (acc.)	mm	✓	✓	✓	✓	✓	17/01/2019	12/02/2019
CLCT	Total cloud cover	%	✓	-	✓	-	✓	18/12/2018	08/01/2019

Two of the variables listed in Table 5.1 require additional explanation, namely, the precipitation rate and sum. The precipitation sum refers to precipitation in the form of rain, snow, hail and graupel accumulated from the initialization time of the model to the end of the lead time. Hence, the corresponding time period is extending and the forecast values are monotonically increasing in the temporal component. In contrast, the precipitation rate refers to the hourly average between two lead times, in particular, the lead time of interest and its predecessor, which result in time periods with a length of 1, 3, 6 or 12 hours. The variable is no direct output of the model, but instead calculated from the precipitation sum. Further, the mean sea level pressure is calculated based on the temperature at 2 m and the soil pressure. Besides that, all variables are direct model output. We refer to Zängl et al. (2015) for detailed descriptions of the ensemble predictions.

Since the portal was launched, the ensemble forecasts on display have been archived at IMK-TRO. However, observations corresponding to the ensemble forecasts were not included and needed to be acquired separately in order to develop the statistical postprocessing models. For the German cities, we obtained the desired observational data for weather stations located near Karlsruhe (KA), Berlin (BE), Hamburg (HH), Mainz (MZ) and Munich (MU) from the Open Data Server of DWD.[5] The following DWD stations have been selected: Berlin-Tempelhof for BE (station ID: 433), Hamburg-Fuhlsbüttel for HH (station ID: 1975), Rheinstetten for KA (station ID: 4177), Mainz-Lerchenberg (ZDF) for MZ (station ID: 3137) and München-Stadt for MU (station ID: 3379). When multiple stations were available for one location, we decided to use the station that includes most of the meteorological variables required. Still, it was not possible to find observations for all forecasts, as not all stations report all of the meteorological variables (Table 5.1). For detailed descriptions on the observations, we refer to Becker and Behrens (2012).

When the idea to implement postprocessing on the portal arose in late 2019, the archive

[5] Although Offenbach is another German city displayed on the portal, it is not included in the postprocessing, as it was added at a later stage of the project. Further, we decided to exclude the other European cities, as observations could not be accessed via DWD.

of the ensemble forecasts included data of almost a year. Table 5.1 shows the first available initialization times in the dataset ending in February 2019 with the integration of the ICON-Global forecasts of precipitation. In early 2020, we began with the implementation of the postprocessing methods and started to regularly retrieve observational data from DWD.[6] On the other end, the last initialization time that was acquired with the last update of the dataset is 7 November 2022.[7]

Training and test data selection

The parameters of the postprocessing models are estimated with the help of ensemble forecasts and corresponding observations from a training dataset, where several options in terms of both spatial and temporal composition can be considered. From the spatial point of view, there are two traditional approaches: local and global selection (Thorarinsdottir and Gneiting, 2010). In the local approach, the parameters of the predictive distribution for a given location are estimated using only data from that particular location, resulting in different parameter estimates for the different locations. In order to ensure numerical stability of the estimation process, local modeling requires long time periods for training, which is the major disadvantage of this approach. As it addresses the location-specific forecast error characteristics, it often results in better forecast skill than global estimation, where training data of the whole ensemble domain is used and all locations share the same set of parameters. Comparing, e.g., the verification rank histograms of the ensemble forecasts for KA and MU displayed in Figures 5.2 and 5.3, we find that the error characteristics differ for each location drastically, such as in case of the mean sea level pressure. Hence, we train separate models for each location.

In the dataset, ensemble predictions of multiple lead times are available and the forecasts are initialized by the NWP model at four different times of the day. These are treated separately when estimating model parameters, i.e., a separate postprocessing model is estimated for each lead time and each initialization hour, based on training datasets comprised of data from those lead times and initialization hours only. Thereby, we aim to account for changes in the forecast error characteristics of the raw ensemble predictions over multiple lead times (Figure 5.4), and for potential diurnal effects by ensuring that the training data covers the same time of day of the observation.

Regarding the temporal composition, the standard approach in EMOS modeling is the use of rolling training periods, where training data consists of forecasts and observations for the n calendar days preceding the target date of interest. Rolling training periods can be flexibly applied to smaller datasets and enable models to adapt to changes in meteorological conditions or the underlying NWP system. An alternative approach is to utilize all available data by considering expanding training periods, motivated by studies suggesting that using

[6] The observations are obtained via the `rdwd` package (Boessenkool, 2021). We are updating the data via the 'recent'-subdirectories, which contain observations from the last 500 days up to the current day.

[7] Due to an outage of the KIT data storage, which IMK-TRO is using to archive the ensemble forecast data, no further updates had been possible.

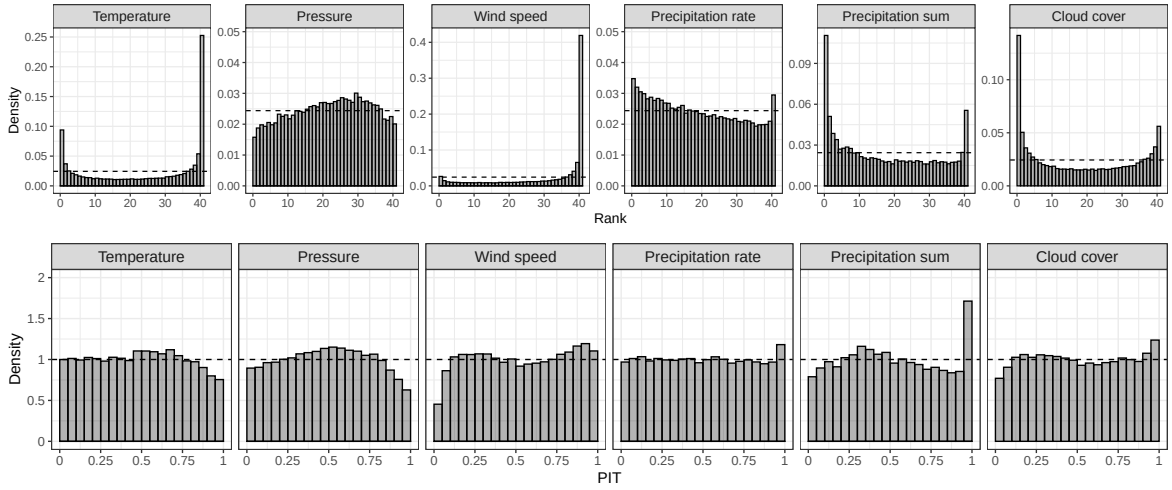

Figure 5.2: Verification rank histograms of the ensemble forecasts (top) and PIT histograms of the postprocessed forecasts (bottom) for the meteorological variables at the station in KA for lead times within two days.

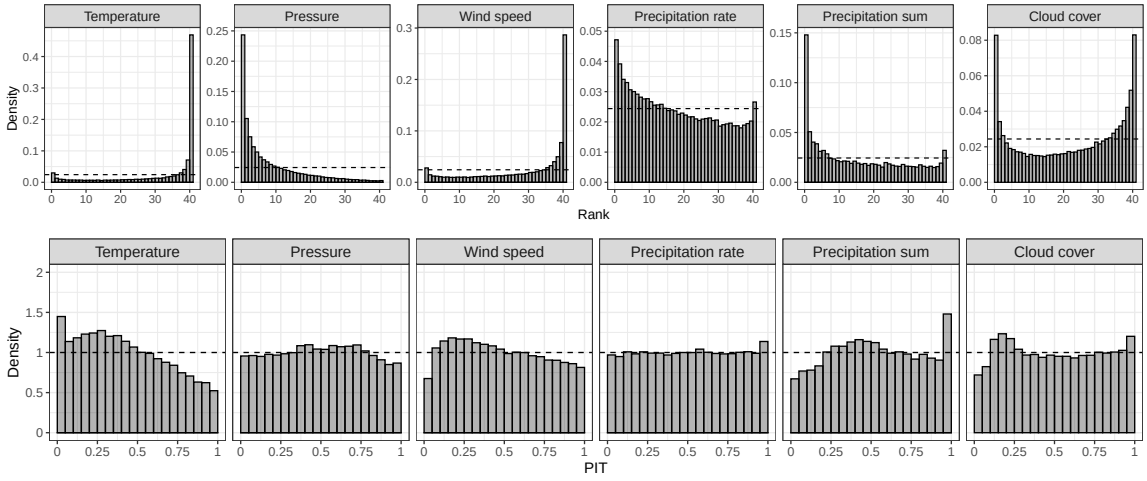

Figure 5.3: Verification rank histograms of the ensemble forecasts (top) and PIT histograms of the postprocessed forecasts (bottom) for the meteorological variables at the station in MU for lead times within two days.

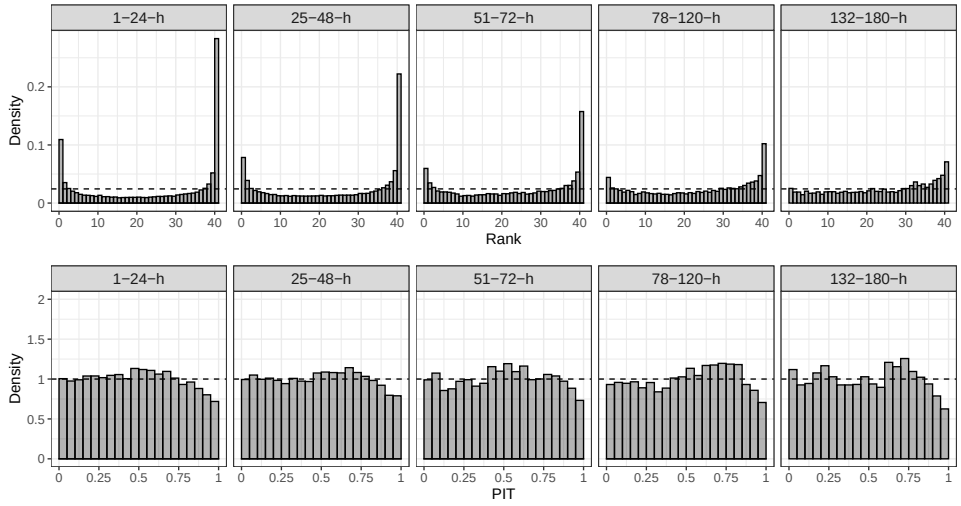

Figure 5.4: Verification rank histograms of the ensemble forecasts (top) and PIT histograms of the postprocessed forecasts (bottom) for temperature at the station in KA for different lead time periods.

long archives of training data irrespective of potential NWP model changes during that period often show superior performance (Lang et al., 2020). Regularly extending training sets may for example be relevant in operational implementations where data archives might be built up and expanded over time.

Here, we faced this decision developing the postprocessing models with less than two years of data. In this case, fixed training periods result in test periods that are too short and, with less than two years, the periods do not cover the same part of the yearly cycle, while CV does not mimic the operational setting for the implementation on the website. On the other side, rolling training periods require frequent updates of the training period in real-time when used operationally on the portal, which increases the complexity of the implementation and is more error prone when data or updates are missing. Further, some variables like precipitation rate or cloud cover require large training sets and are not optimally suited for the use of rolling training periods. Instead, we decided to use the alternative approach presented above, namely a monthly extending training period, that is, for each test sample, we use all data available until the end of the month preceding the day of interest for training of the underlying model. In an operational context, this means that we update the training set and train the postprocessing models only once a month.

For the evaluation of the postprocessed forecasts, we define the test period as the year preceding the last initialization time and use the rest of the data as initial training period. Then, the training set is updated monthly, that is, each time a test sample starts a new month. Precisely, we are using the period from 8 November 2021 to 7 November 2022 for testing and the period from the start of the archive in late 2018 or early 2019 (dependent on the variable; Table 5.1) to 7 November 2021 as initial training period.

5.1.2 Model configurations

As mentioned above, we apply the EMOS approach presented in Section 4.2.1 for postprocessing, except for cloud cover.

EMOS models

The first important choice for EMOS is the forecast distribution. Based on the discussion in Section 4.2.1, we have decided on the parametric distributions given in Table 5.2. Note that the zero-truncated and zero-censored logistic distribution are described in Section 2.3.1. For the parameterizations, which are also given in Table 5.2, we use the notation introduced in Section 4.2, in particular, $x = (x_1, \ldots, x_{40})$ here denotes the ensemble forecast of the variable of interest. The parameterization of the precipitation variables includes a summary statistic that has not been introduced yet, namely, the *fraction of zero-ensemble members* $p_0(x)$, which is in general defined for an ensemble x of size m via

$$p_0(x) := \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\{x_i = 0\}. \tag{5.1}$$

The fraction of zero-ensemble members was included due to a large number of zero-ensemble-observation pairs, especially for the precipitation rate. Zero-ensemble forecasts also result in standard deviations of zero, which cause numerical problems as argument of the logarithm. Therefore, we shift standard deviations of zero on a small threshold, namely 10^{-4}. Because the precipitation rate includes the largest amount of zeros, we changed the maximum number of iterations in the optimization from the default value of 100 steps to 10, following the idea of Scheuerer (2014). The parameterization of the scale parameter for the precipitation and wind speed follows Jordan et al. (2019), while that of the location parameter for precipitation was found experimenting with examples from Vannitsem et al. (2018, p. 59), Baran and Lerch (2016) and Jordan et al. (2019). Further, we restrict the wind speed models to positive location parameters $\mu > 0$. As described in Section 2.3.1, the mode of the truncated logistic distribution is given by $\max(\mu, 0)$ and the reduction in flexibility can thus be seen as a restriction to positive modes. Given that negative location parameters tend to be estimated only for wind speeds of low-intensity, which are in general of little interest, and that we noticed no effect on the predictive performance, the restriction of the parameter space can be considered to be negligible.

The EMOS parameters are determined using optimum score estimation based on the CRPS, where we use the L-BFGS-B method (Byrd et al., 1995) in the gradient-based optimization described in Section 4.2.1.[8]

[8] L-BFGS-B refers to an algorithm that uses a Limited memory BFGS (Broyden–Fletcher–Goldfarb–Shanno) matrix for Bound constrained optimization.

Table 5.2: Overview of the EMOS models used on the KIT-Weather portal. Note that $c, d > 0$ for temperature and mean sea level pressure, otherwise all EMOS parameters are real-valued.

Met. var.	Distribution	Location	Scale
T_2M	Gaussian	$a + b\overline{x}$	$\sqrt{c + ds^2(x)}$
MSLP	Gaussian	$a + b\overline{x}$	$\sqrt{c + ds^2(x)}$
WIND_10M	Zero-truncated logistic	$\exp(a + b\log(\overline{x}))$	$\exp(c + d\log(s(x)))$
PREC_RATE	Zero-censored logistic	$a + b_0 p_0(x) + b_1\overline{x}$	$\exp(c + d\log(s(x)))$
PREC_SUM	Zero-censored logistic	$a + b_0 p_0(x) + b_1\overline{x}$	$\exp(c + d\log(s(x)))$

Cloud cover

Unlike the other variables, the ensemble forecasts of cloud cover are not postprocessed using EMOS. Although the ICON-EPS generates continuous forecasts, which take values from 0 to 1 representing the degree of cloudiness in percentage, total cloud cover is observed in classes from 0 to 8 corresponding to the values $\{0, 0.1, 0.25, 0.4, 0.5, 0.6, 0.75, 0.9, 1\}$ (Hemri et al., 2016, Table A1). In order to account for the discrete observations, we use a postprocessing method tailored to cloud cover that forecasts class probabilities, namely the *proportional odds logistic regression* (POLR) model presented in Hemri et al. (2016). The POLR approach is especially suited for ordinal targets and performs a multinomial logistic regression assuming proportional odds, that is, predictor coefficients that are constant for all classes. This results in a more parsimonious model, as only one coefficient for each predictor needs to be estimated, in addition to one intercept per class.

For a mathematical formulation, we define the classes $c_k = k$ together with the accumulated probabilities $\pi_k = \mathbb{Q}(Y \leq c_k)$ for $k = 0, \ldots, 8$. Then, Hemri et al. (2016, equation (4)) define the POLR model via

$$\operatorname{logit}(\pi_k) = \log \frac{\pi_k}{1 - \pi_k} = \theta_k - \boldsymbol{\beta}^T \boldsymbol{x}, \quad k = 0, \ldots, 8, \tag{5.2}$$

where $\boldsymbol{x} \in \mathbb{R}^p$ denotes the predictor vector, $\boldsymbol{\beta} \in \mathbb{R}^p$ the (joint) regression coefficient vector and $\theta_k \in \mathbb{R}$ the (classwise) intercepts with $\theta_0 < \cdots < \theta_8$. The class probabilities p_k are the increments of the accumulated probabilities, i.e., $p_0 = \pi_0$ and $p_k = \pi_k - \pi_{k-1}$ for $k = 1, \ldots, 8$.

Next to the ensemble mean and variance, we are using the fraction of 100%-forecasts as an additional predictor following the suggestions of Hemri et al. (2016). It improves the overall performance and is calculated analogous to the fraction of zero-ensemble members in equation (5.1), which did not show improvements in forecast performance. As for EMOS, we estimate a separate model based on a monthly extending train period for each location, lead time and initialization hour.

For the meteograms displayed on the website, we generate a 40-member ensemble by randomly drawing from a piecewise uniform distribution, where each bin corresponds to one of the nine classes of cloud coverage defined in Hemri et al. (2016, Table A1). While we generate continuous forecasts for the portal, we evaluate the forecasts based on the discrete observations and rely on the formula used in Hemri et al. (2016, equation (A3)) for the calculation of the CRPS.

5.1.3 RESULTS

In this section, we will evaluate the performance of the postprocessing methods. To illustrate the effects of statistical postprocessing, we first need to investigate the behavior of the ensemble forecasts.

PREDICTIVE PERFORMANCE OF THE ICON-EPS FORECASTS

The verification rank histograms in Figure 5.2, which summarize the forecasts for KA for lead times of up to two days, show different types of miscalibration. The forecasts for wind speed, temperature, cloud cover, precipitation sum and rate are underdispersed with the degree of miscalibration decreasing in that order. In contrast to the typical underdispersion, the ensemble forecasts of mean sea level pressure are overdispersed. Next to dispersion, we observe a strong negative bias for the temperature and wind speed forecasts, meaning that the ensemble frequently underforecasts the actually observed value, and a positive bias for precipitation sum and cloud cover. The rank histograms of precipitation rate and mean sea level pressure also exhibit a small bias.

In general, we observe underdispersion and biases for all stations. The magnitude of the systematic errors varies however, e.g., the ensemble forecasts exhibit a strong bias for the station in MU (Figure 5.3). Hence, the grid point and station selected for MU are not as representative of each other as for the other stations. The systematic errors do not only change for the different stations, but also for the lead times. The rank histograms of the temperature forecasts for KA in Figure 5.4 exemplify that the longer the lead time, the smaller the degree of underdispersion. This can be explained by the fact that the ensemble range increases with the lead time, as illustrated by Figure 5.5. The growth of the ensemble range is in line with the intuition that the forecast uncertainty increases with the lead time due to a lower predictability.

Even though the degree of miscalibration decreases, the predictive performance measured in terms of the CRPS becomes worse, as Tables 5.3 and 5.4 demonstrate. Over all stations, the CRPS of the precipitation sum decreases by 898% from lead times within the first day to lead times longer than five days, for mean sea level pressure by 272%, for temperature by 81%, for cloud cover by 39% and for wind speed by 29%, while it does not change for the precipitation rate. Regarding the extreme decrease in predictive performance for the precipitation sum, we

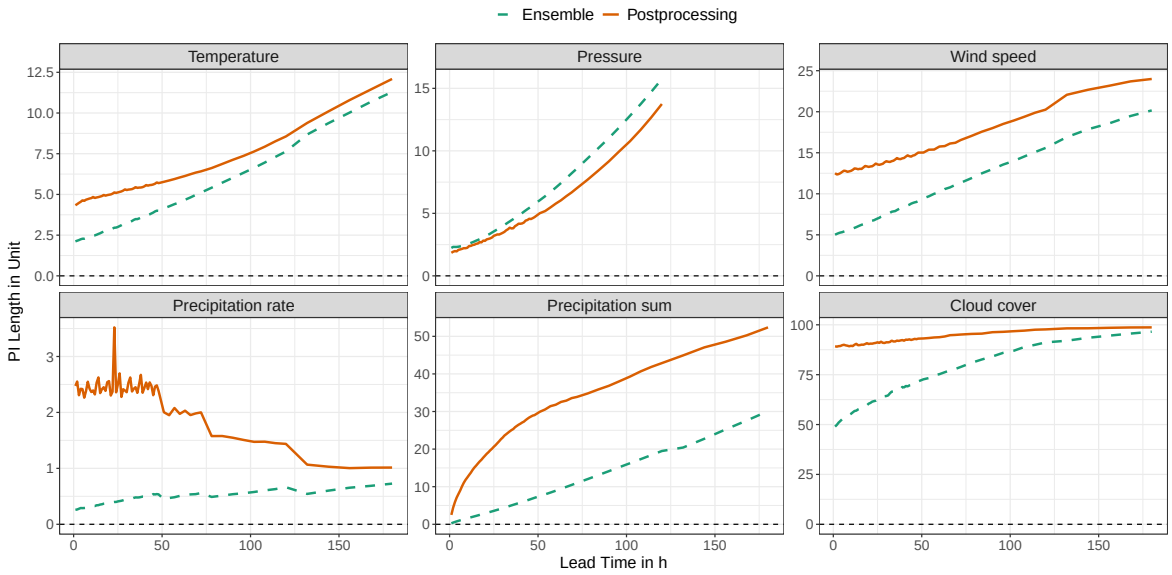

Figure 5.5: Ensemble range (dashed) and length of 95.12% PIs of the postprocessed forecasts (solid) for the meteorological variables dependent on the lead time in hours averaged over all stations.

have to take into account that the period covered by the forecast is dependent on the lead time. Therefore, more zero-ensemble-observation pairs, which reduce the mean CRPS, occur for smaller lead times. Similar arguments hold for the precipitation rate, which has a small CRPS of almost zero for all lead times. This can be explained by the fact that the fraction of zero-ensemble-observation pairs is even larger.[9]

Evaluation of the postprocessed forecasts

Overall, postprocessing improves the predictive performance with respect to the raw ensemble for all lead times. Comparing the verification rank histograms that we used to evaluate the predictive performance of the ensemble forecasts with the PIT histograms of the associated postprocessed forecasts in Figure 5.2, we find that postprocessing corrects the miscalibration and results in forecasts that are much better calibrated exhibiting histograms that are almost flat. However, we still find deviations from uniformity. For the mean sea level pressure, the forecasts are still overdispersed, and for the precipitation sum, the last bin is the most frequent meaning that we underforecast the accumulated precipitation sum, here almost twice as much as expected if the forecasts were calibrated. In case of the precipitation rate, the last bin is also the most frequent, but to a much smaller extent, which might again be due to the large amount of zeros in the data. For temperature and wind speed, the histograms exhibit peculiar shapes that stem from the predictive distribution being constrained to the shape of

[9]Within the evaluation period, the fraction of zero-ensemble-observation pairs is 46%, the fraction of zero-observations is 88% and the average ensemble mean (median) is 0.07 (0.001) mm/h.

Table 5.3: Mean CRPS of the postprocessed (ensemble) forecasts for the meteorological variables for lead times within two days.

Lead times		1–24 hours				25–48 hours			
Init. hour		00 UTC	06 UTC	12 UTC	18 UTC	00 UTC	06 UTC	12 UTC	18 UTC
T_2M	KA	0.80 (0.92)	0.80 (0.93)	0.80 (0.96)	0.81 (0.94)	0.88 (0.99)	0.89 (1.00)	0.88 (0.99)	0.89 (1.00)
	BE	0.59 (0.74)	0.60 (0.78)	0.58 (0.76)	0.59 (0.74)	0.71 (0.83)	0.72 (0.85)	0.70 (0.83)	0.70 (0.83)
	HH	0.63 (0.71)	0.64 (0.73)	0.63 (0.71)	0.65 (0.74)	0.74 (0.81)	0.75 (0.82)	0.75 (0.81)	0.75 (0.81)
	MZ	0.63 (0.80)	0.63 (0.83)	0.62 (0.85)	0.64 (0.83)	0.72 (0.88)	0.72 (0.90)	0.72 (0.90)	0.73 (0.91)
	MU	0.69 (1.09)	0.69 (1.15)	0.68 (1.15)	0.70 (1.07)	0.80 (1.17)	0.80 (1.20)	0.79 (1.19)	0.81 (1.16)
MSLP	KA	0.31 (0.32)	0.32 (0.33)	0.30 (0.32)	0.30 (0.31)	0.49 (0.50)	0.49 (0.50)	0.48 (0.49)	0.49 (0.50)
	BE	0.28 (0.29)	0.28 (0.29)	0.27 (0.29)	0.28 (0.29)	0.52 (0.53)	0.51 (0.52)	0.50 (0.52)	0.52 (0.53)
	HH	0.27 (0.29)	0.27 (0.29)	0.27 (0.29)	0.27 (0.29)	0.50 (0.52)	0.50 (0.53)	0.49 (0.51)	0.50 (0.52)
	MU	0.37 (0.72)	0.37 (0.67)	0.35 (0.69)	0.36 (0.79)	0.55 (0.83)	0.55 (0.81)	0.55 (0.85)	0.54 (0.84)
WIND_10M	KA	1.93 (3.06)	1.95 (3.06)	1.94 (3.10)	1.94 (3.07)	2.11 (3.11)	2.13 (3.08)	2.10 (3.06)	2.10 (3.05)
	BE	1.71 (2.85)	1.73 (2.87)	1.74 (2.86)	1.72 (2.81)	1.90 (2.85)	1.91 (2.87)	1.91 (2.87)	1.89 (2.82)
	HH	1.86 (3.62)	1.86 (3.64)	1.84 (3.58)	1.84 (3.50)	2.11 (3.66)	2.09 (3.65)	2.08 (3.61)	2.09 (3.55)
	MU	1.37 (1.87)	1.37 (1.88)	1.35 (1.87)	1.36 (1.85)	1.49 (1.95)	1.49 (1.95)	1.50 (1.97)	1.50 (1.96)
PREC_RATE	KA	0.06 (0.06)	0.07 (0.07)	0.06 (0.07)	0.06 (0.06)	0.07 (0.07)	0.07 (0.07)	0.07 (0.07)	0.07 (0.07)
	BE	0.04 (0.04)	0.04 (0.04)	0.04 (0.04)	0.04 (0.04)	0.04 (0.04)	0.04 (0.04)	0.04 (0.04)	0.04 (0.04)
	HH	0.06 (0.06)	0.06 (0.06)	0.06 (0.06)	0.06 (0.06)	0.06 (0.06)	0.06 (0.06)	0.06 (0.06)	0.06 (0.06)
	MZ	0.04 (0.04)	0.04 (0.05)	0.05 (0.05)	0.04 (0.04)	0.05 (0.05)	0.05 (0.05)	0.05 (0.05)	0.05 (0.05)
	MU	0.07 (0.08)	0.07 (0.09)	0.07 (0.08)	0.07 (0.08)	0.07 (0.08)	0.08 (0.09)	0.08 (0.09)	0.08 (0.09)
PREC_SUM	KA	0.41 (0.44)	0.52 (0.56)	0.52 (0.57)	0.47 (0.50)	1.30 (1.38)	1.43 (1.52)	1.41 (1.52)	1.37 (1.45)
	BE	0.26 (0.29)	0.32 (0.36)	0.32 (0.35)	0.24 (0.27)	0.72 (0.80)	0.77 (0.86)	0.75 (0.82)	0.70 (0.79)
	HH	0.45 (0.47)	0.52 (0.54)	0.57 (0.60)	0.40 (0.41)	1.19 (1.24)	1.26 (1.32)	1.31 (1.37)	1.14 (1.17)
	MZ	0.35 (0.35)	0.34 (0.36)	0.35 (0.36)	0.34 (0.35)	0.96 (0.97)	0.93 (0.94)	0.96 (0.97)	0.97 (0.98)
	MU	0.46 (0.58)	0.57 (0.68)	0.54 (0.63)	0.52 (0.61)	1.26 (1.54)	1.40 (1.68)	1.31 (1.56)	1.31 (1.55)
CLCT	KA	13.73 (15.62)	13.76 (15.72)	13.67 (15.75)	13.66 (15.65)	14.42 (16.05)	14.61 (16.19)	14.69 (16.42)	14.62 (16.27)
	HH	11.20 (11.38)	11.24 (11.49)	11.15 (11.44)	11.21 (11.45)	12.00 (12.15)	11.82 (11.93)	11.72 (11.87)	11.82 (11.99)
	MU	11.49 (13.15)	11.51 (13.16)	11.46 (13.16)	11.46 (13.05)	12.25 (13.67)	12.49 (13.96)	12.63 (14.20)	12.44 (13.84)

the chosen parametric family, e.g., a lower tail that is too heavy in case of the wind speed.

The postprocessed forecasts for MU still offer large improvements over the ensemble, but result in less well-calibrated forecasts than for KA, as the postprocessed temperature forecast is now biased in the opposite direction as the ensemble (Figure 5.3). The mean sea level pressure forecasts are however less dispersed than those for KA and therefore better calibrated. The comparison of the degree of calibration over the lead times is striking, since the calibration of the postprocessed temperature forecasts for KA does not seem to be dependent on the lead time, as the histograms have a similar shape for all lead times (Figure 5.4).

The improvement becomes apparent when comparing the CRPS values of the ensemble and postprocessed forecasts for lead times within two days (Table 5.3). In all of the 208 combinations of location, lead time and initialization hour, the table does not show one decrease in the CRPS when rounded to the second digit.[10] Instead, the CRPS improves throughout all variables besides the precipitation rate. The CRPS values for lead times from two to six and a half days exhibit the same pattern (Table 5.4). Further, we note that the CRPS values have a different magnitude for each variable, which impedes comparability over variables.

By showing the CRPSS, Figure 5.6 provides more insight in the improvement obtained

[10]The CRPS actually decreases in the fifth digit for two cases of precipitation rate in HH.

Table 5.4: Mean CRPS of the postprocessed (ensemble) forecasts for the meteorological variables for lead times over two days.

Lead times		51–72 hours				78–120 hours				132–180 hours	
Init. hour		00 UTC	06 UTC	12 UTC	18 UTC	00 UTC	06 UTC	12 UTC	18 UTC	00 UTC	12 UTC
T_2M	KA	0.94 (1.02)	0.96 (1.04)	0.96 (1.06)	0.96 (1.03)	1.11 (1.17)	1.12 (1.18)	1.14 (1.22)	1.14 (1.20)	1.48 (1.55)	1.53 (1.61)
	BE	0.79 (0.89)	0.81 (0.90)	0.80 (0.90)	0.79 (0.89)	0.97 (1.05)	0.98 (1.04)	0.96 (1.04)	0.97 (1.05)	1.40 (1.47)	1.45 (1.50)
	HH	0.83 (0.88)	0.86 (0.90)	0.86 (0.90)	0.86 (0.90)	1.01 (1.04)	1.01 (1.04)	1.02 (1.06)	1.03 (1.06)	1.40 (1.44)	1.44 (1.47)
	MZ	0.78 (0.91)	0.80 (0.94)	0.77 (0.94)	0.78 (0.93)	0.95 (1.05)	0.96 (1.07)	0.95 (1.08)	0.96 (1.08)	1.37 (1.57)	1.41 (1.62)
	MU	0.89 (1.19)	0.89 (1.24)	0.87 (1.23)	0.88 (1.17)	1.01 (1.24)	1.02 (1.27)	0.98 (1.27)	1.01 (1.23)	1.46 (1.59)	1.50 (1.65)
MSLP	KA	0.74 (0.74)	0.75 (0.76)	0.74 (0.74)	0.76 (0.77)	1.25 (1.27)	1.30 (1.31)	1.29 (1.32)	1.29 (1.31)	-	-
	BE	0.80 (0.82)	0.82 (0.84)	0.81 (0.82)	0.82 (0.84)	1.45 (1.47)	1.49 (1.52)	1.48 (1.51)	1.50 (1.52)	-	-
	HH	0.83 (0.85)	0.84 (0.87)	0.84 (0.86)	0.86 (0.88)	1.57 (1.58)	1.58 (1.61)	1.60 (1.62)	1.62 (1.63)	-	-
	MU	0.78 (0.97)	0.78 (0.97)	0.79 (0.99)	0.80 (0.98)	1.27 (1.35)	1.33 (1.39)	1.33 (1.40)	1.32 (1.38)	-	-
WIND_10M	KA	2.36 (3.18)	2.35 (3.14)	2.34 (3.15)	2.33 (3.02)	2.85 (3.57)	2.83 (3.56)	2.82 (3.53)	2.87 (3.49)	3.55 (4.39)	3.54 (4.36)
	BE	2.11 (2.94)	2.13 (3.01)	2.14 (2.95)	2.11 (2.88)	2.53 (3.14)	2.56 (3.18)	2.57 (3.17)	2.55 (3.11)	3.18 (3.55)	3.23 (3.62)
	HH	2.41 (3.82)	2.43 (3.84)	2.38 (3.63)	2.32 (3.48)	2.95 (4.06)	2.98 (4.11)	2.94 (3.94)	2.93 (3.83)	3.67 (3.96)	3.70 (4.01)
	MU	1.65 (2.06)	1.65 (2.09)	1.64 (2.08)	1.62 (2.00)	1.92 (2.30)	1.92 (2.32)	1.90 (2.29)	1.92 (2.26)	2.43 (2.63)	2.39 (2.58)
PREC_RATE	KA	0.06 (0.07)	0.07 (0.07)	0.07 (0.07)	0.07 (0.07)	0.07 (0.07)	0.07 (0.07)	0.07 (0.07)	0.07 (0.07)	0.07 (0.08)	0.07 (0.08)
	BE	0.04 (0.04)	0.04 (0.04)	0.04 (0.04)	0.03 (0.04)	0.04 (0.04)	0.04 (0.04)	0.04 (0.04)	0.04 (0.04)	0.04 (0.04)	0.04 (0.04)
	HH	0.06 (0.06)	0.06 (0.06)	0.06 (0.06)	0.06 (0.06)	0.06 (0.06)	0.06 (0.06)	0.06 (0.06)	0.06 (0.06)	0.06 (0.06)	0.06 (0.06)
	MZ	0.05 (0.05)	0.05 (0.05)	0.05 (0.05)	0.05 (0.05)	0.05 (0.05)	0.05 (0.05)	0.05 (0.05)	0.05 (0.05)	0.05 (0.05)	0.05 (0.05)
	MU	0.07 (0.08)	0.08 (0.09)	0.07 (0.08)	0.07 (0.08)	0.07 (0.07)	0.08 (0.08)	0.07 (0.08)	0.07 (0.08)	0.07 (0.07)	0.07 (0.07)
PREC_SUM	KA	2.17 (2.30)	2.30 (2.44)	2.27 (2.43)	2.28 (2.40)	3.50 (3.70)	3.68 (3.89)	3.65 (3.85)	3.71 (3.88)	6.02 (6.13)	6.05 (6.25)
	BE	1.16 (1.30)	1.22 (1.36)	1.20 (1.33)	1.17 (1.33)	1.86 (2.07)	1.97 (2.17)	1.96 (2.21)	1.90 (2.12)	2.84 (2.99)	2.86 (3.06)
	HH	1.90 (1.98)	1.99 (2.08)	2.02 (2.10)	1.83 (1.89)	2.97 (3.10)	3.00 (3.13)	3.08 (3.19)	2.92 (3.01)	4.55 (4.85)	4.54 (4.80)
	MZ	1.57 (1.58)	1.58 (1.59)	1.59 (1.60)	1.59 (1.59)	2.41 (2.37)	2.41 (2.41)	2.45 (2.46)	2.45 (2.45)	3.79 (3.79)	3.81 (3.82)
	MU	2.01 (2.40)	2.20 (2.65)	2.10 (2.49)	2.07 (2.42)	3.10 (3.65)	3.33 (3.91)	3.25 (3.77)	3.20 (3.67)	4.76 (5.25)	4.77 (5.31)
CLCT	KA	15.33 (16.59)	15.55 (16.77)	15.27 (16.49)	15.53 (16.74)	17.25 (18.23)	17.09 (18.09)	17.24 (18.31)	17.38 (18.41)	19.32 (20.12)	19.58 (20.62)
	HH	12.79 (12.85)	13.04 (13.11)	13.22 (13.24)	12.92 (13.01)	14.76 (14.79)	14.62 (14.65)	15.02 (15.14)	14.74 (14.81)	17.02 (17.19)	16.96 (17.15)
	MU	13.30 (14.67)	13.58 (14.91)	13.50 (14.68)	13.33 (14.41)	14.94 (15.83)	14.88 (15.76)	14.70 (15.66)	14.49 (15.37)	17.27 (17.80)	17.43 (18.03)

by postprocessing, as it allows for comparisons over the meteorological variables. First, we observe that postprocessing improves the predictive performance with respect to the ensemble forecasts for all cases besides precipitation rate and sum in MZ for lead times larger than 78 hours and some lead times of precipitation rate in HH. As each station exhibits different error characteristics with systematic errors of varying magnitude, the skill is also varying for the different stations. One example are the mean sea level pressure forecasts that are strongly biased for MU and result in a skill that is orders of magnitude larger than for the other locations. Further, the skill decreases as the lead time increases. This might be a result of the fact that the degree of underdispersion of the ensemble forecasts decreases with the lead time, thus the degree of systematic errors that require correction (Figure 5.4). The change of the underlying ICON model also has an effect on the skill curves, as can be seen in the breaks, e.g., for temperature in MZ or wind speed in HH. The reason is that together with the model the resolution changes and therefore different error characteristics result in different skill scores. Again, the precipitation rate is a special case. Here, the skill curve is fluctuating due to small values of the CRPS. Note that we show a smoothed version of the CRPSS that is based on the mean CRPS of the period from 5 hours before and after the lead time of interest. This was done to eliminate the diurnal cycle, which is present although four different initialization hours are taken into account (Figure 5.7).

In addition, one forecast bust occurs for the precipitation rate in BE for a lead time of 23 hours. The term forecast bust is ambiguous and generally refers to (a set of) degenerate

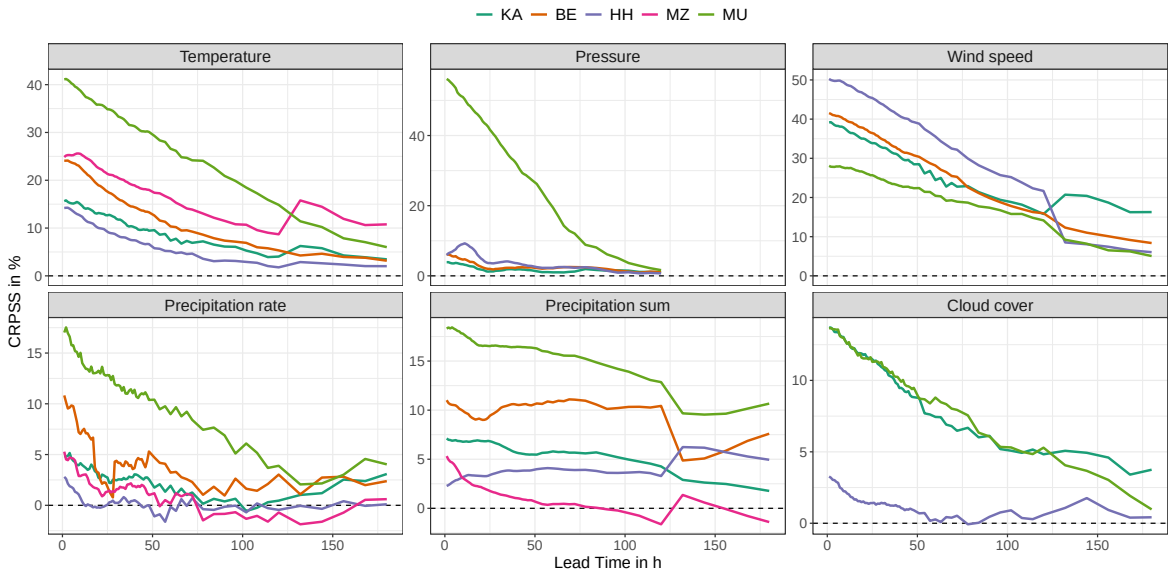

Figure 5.6: Smoothed CRPSS of the postprocessed forecasts with respect to the ensemble forecasts for the meteorological variables at the different stations dependent on the lead time in hours. The mean CRPS underlying the skill score is calculated as a rolling mean of the corresponding lead time ±5 hours.

forecasts, which includes forecasts that are physically inconsistent, unrealistic and/or perform exceptionally poor. Due to overfitting on the training data, one ensemble forecast with a large standard deviation resulted in an extreme scale parameter.[11] The forecast bust is still visible in Figure 5.6, as the skill curve drops down for BE and lead times from 18 to 28 hours. In Figures 5.7 and 5.8, the bust becomes apparent in the negative peak of the skill curve and the peak of the PI length.

The sharpness of the postprocessed forecasts is evaluated based on the length of 95.12% PIs, which correspond to the nominal coverage of a calibrated 40-member ensemble. Figure 5.5 shows the PI length averaged over all stations and initialization hours in comparison with the ensemble range. For the underdispersed variables temperature, wind speed, precipitation rate and precipitation sum, postprocessing increases the PI length with respect to the ensemble range. Contrarily, the PI length decreases for the overdispersed ensemble forecasts of mean sea level pressure. Again, precipitation rate is the most unstable variable, for which we interestingly observe drops in the ensemble range for lead times with a change in temporal resolution. The precipitation rate describes the hourly amount of precipitation over a given

[11] The large amount of (almost) zero-ensemble forecasts in the data results in ensemble standard deviations close to zero. The EMOS parameters are then fitted based on these standard deviations and yield overly large scale parameters when the ensemble variance actually becomes large. In case of the forecast bust, no precipitation was observed while a large standard deviation resulted in an extreme scale parameter, where the postprocessed forecast had a CRPS of 18.7 mm/h and the ensemble forecast a CRPS of only 0.6 mm/h. Comparing the values with the entries of Table 5.3, we can explain the strong effect of one sample on the forecast skill. Note that such cases can be avoided by checking and constraining the parameter values.

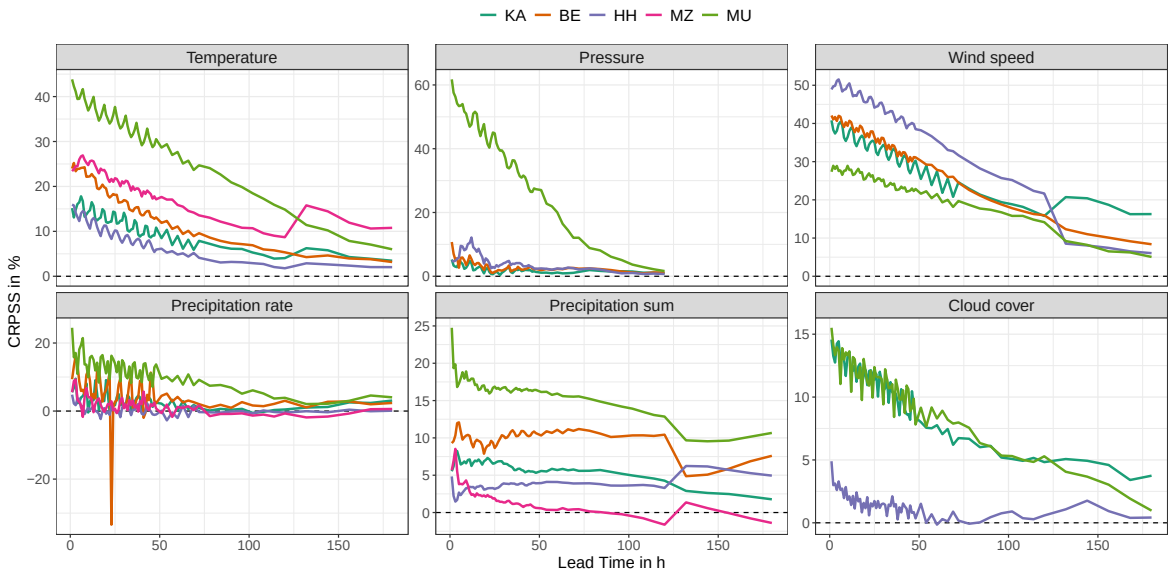

Figure 5.7: CRPSS of the postprocessed forecasts with respect to the ensemble forecasts for the meteorological variables at the different stations dependent on the lead time in hours.

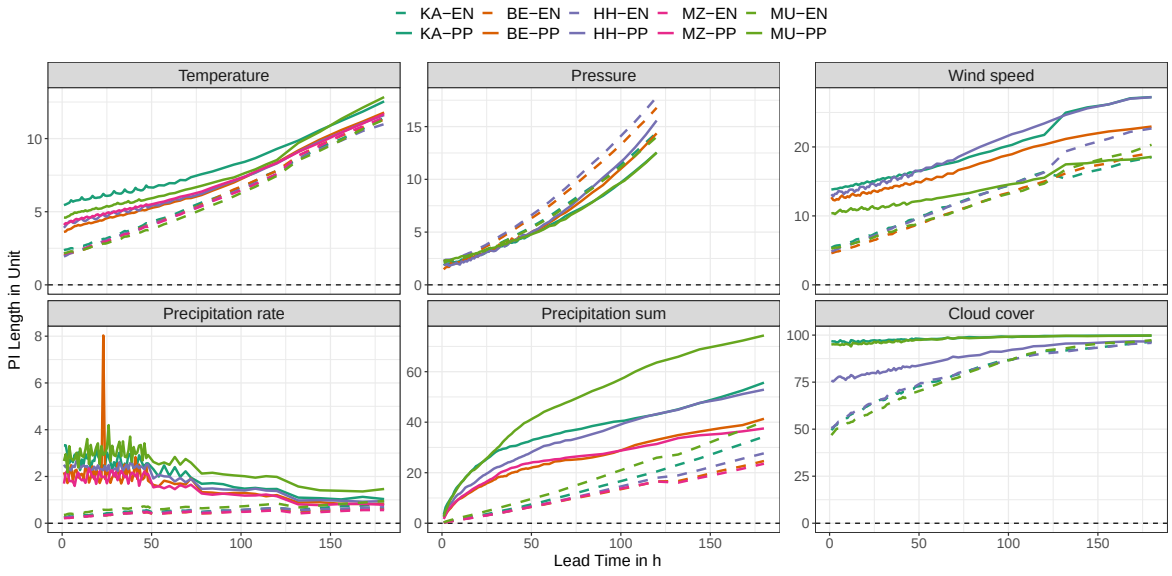

Figure 5.8: Mean ensemble range (dashed) and mean length of 95.12% PIs of the postprocessed forecasts (solid) for the meteorological variables at the different stations dependent on the lead time in hours.

reference period, which is dependent on the temporal resolution and therefore increases with the lead time. The longer the period becomes, the smaller the effect of a single precipitation event within that period and hence the smaller the forecast uncertainty. When comparing the PI lengths at different locations, we find smaller differences than for the skill with curves that exhibit the same shape. Note that while the skill is dependent both on the forecast and observation, the PI length is a property of the forecast alone.[12]

CONCLUSIONS

The basic statistical postprocessing approach used in this section corrects the systematic errors of the ensemble forecasts for the various meteorological variables and lead times. Typical types of miscalibration, such as under- and overdispersion, are corrected by adjusting the uncertainty of the predictive distribution. Up to three days, we obtain clear improvements in the CRPS for almost all cases. Not surprisingly, precipitation has proven to be a difficult task for postprocessing, as noted in other studies (e.g., Scheuerer, 2014).

In the context of dissemination, one major shortcoming is that univariate postprocessing may lead to physical inconsistencies. This becomes apparent for forecasts of the precipitation sum, which is accumulated over the period from the initialization to the end of the lead time and must therefore be nondecreasing as a function of the lead time. However, Figure 5.9 shows an ensemble boxplot meteogram of postprocessed forecasts for the precipitation sum that are physically inconsistent due to a decrease in the accumulated precipitation. Hence, we do not use the model presented in this section to generate postprocessed forecasts of the precipitation sum on the portal, but instead accumulate the postprocessed forecasts of the precipitation rate.[13] This does not only eliminate the inconsistencies within the meteogram of the precipitation sum forecasts but also between the postprocessed versions of the precipitation sum and rate. Further, the generation of the postprocessed forecasts of the precipitation sum mirrors the derivation of the ensemble forecasts of the precipitation rate.

A second physical inconsistency was observed for temperature forecasts for KA during a heat wave in June 2022.[14] Figure 5.9 shows meteograms of the ensemble and postprocessed forecasts of temperature, where both 108 hour forecasts predict the highest temperatures with a median of around 34°C on 19 June. However, while the ensemble results in a skewed boxplot with a lower whisker down below 20°C, the normal distribution underlying the EMOS model enforces a symmetric boxplot with a lower whisker down to 23°C and an upper whisker up to an unrealistically high temperature of almost 45°C. The same pattern can be observed

[12] Although the PI length is not dependent on observations, the EMOS coefficients used to calculate the scale parameter are estimated based on the observations in the training set. Hence, large uncertainties in the observations are reflected in the PI length of the postprocessed forecasts.

[13] At the time the screenshot was taken, the ensemble boxplot meteograms for the precipitation sum were generated based on (directly) postprocessed forecasts using the approach evaluated in this section.

[14] While the inconsistency regarding the precipitation sum was observed shortly after the implementation on the portal, the particular example for temperature was found in an outreach activity when a journalist consulted Prof. Dr. Andreas Fink at IMK-TRO due to the upcoming heat wave.

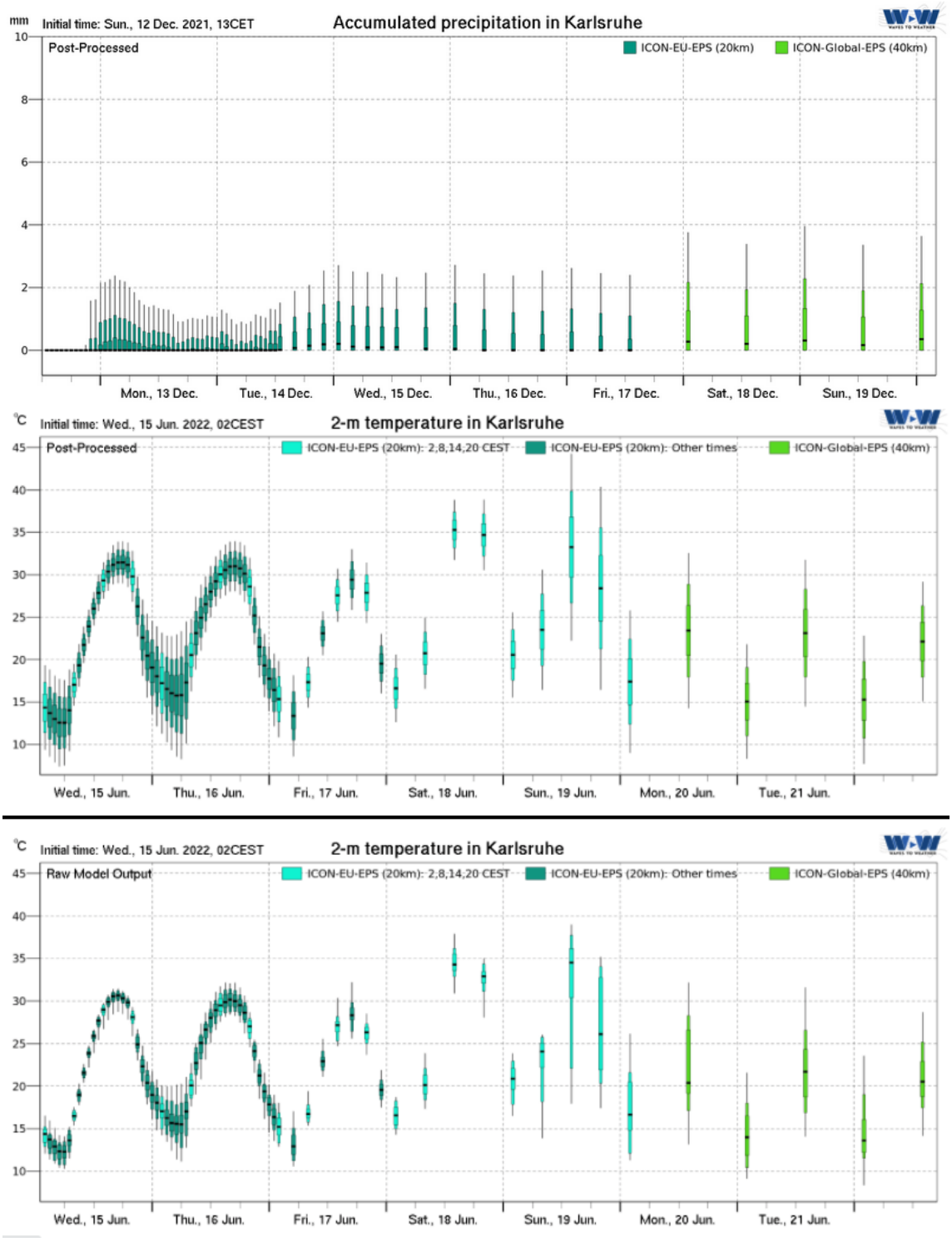

Figure 5.9: Ensemble boxplot meteograms on KIT-Weather of (directly) postprocessed forecasts for precipitation sum in KA (from 14 December 2021; top), and of both postprocessed (middle) and ensemble forecasts (bottom) for temperature in KA (from 15 June 2022). The underlying model runs were initialized at 12 UTC (13 CET) on 12 December 2021, and 00 UTC (02 CEST) on 14 June 2022, respectively.

on a smaller scale for the 102 hour forecast. Two aspects are important for the explanation of this behavior. Due to the implicit assumption of symmetry, which follows from the choice of the normal distribution, the EMOS forecast is not able to generate forecasts that are skewed. Secondly, we use only the ensemble mean and standard deviations as predictor variables, hence the model has no information on the skewness of the ensemble forecasts. However, as pointed out by Gebetsberger et al. (2019), the choice of the response distribution is a much more important factor than the inclusion of additional predictor variables. A similar problem was observed by Gneiting et al. (2023, Figure 1) for forecasts of hourly solar irradiance, where a DRN model based on a (censored) normal distribution was not able to reflect the more realistic shapes of nonparametric IDR and BQN variants. Hence, more flexible forecast distributions such as the BQN forecast might be able to resolve the forecast uncertainty in a physically more realistic way, but require more data and are prone to overfitting in contrast to the EMOS model at hand.

5.2 Solar irradiance forecasting

The KIT-Weather portal also includes forecasts of two solar irradiance variables, which we excluded from the previous section to investigate the topic in more detail. At the begin of the implementation of the postprocessing models, research on statistical postprocessing of ensemble forecasts of solar irradiance variables was scarce. Therefore, we decided to enhance the EMOS models we developed for the KIT-Weather portal towards a case study of larger extent. Here, we highlight the socio-economic relevance of solar irradiance forecasts, introduce a second dataset from Hungary and propose advanced EMOS models that we evaluate on the two datasets.

To reduce emissions of greenhouse gases, a transition towards renewable energy sources such as wind and solar power is imperative (van der Meer et al., 2018). Accurate and reliable forecasts of power generation from those sources are thus becoming increasingly important for integrating volatile power systems into the electrical grid in order to balance demand and supply (Gottwalt et al., 2017; González Ordiano et al., 2020). The literature on energy forecasting has primarily focused on deterministic prediction for the past decades. However, it has now been widely argued that probabilistic forecasting is essential for optimal decision making in planning and operation (Hong and Fan, 2016; van der Meer et al., 2018; Haupt et al., 2019; Hong et al., 2020). For example, Hong et al. (2020) identify probabilistic forecasting with the aim of providing a predictive probability distribution for a future quantity or event in order to quantify forecast uncertainty as one of the most important emerging research topics in their recent review on energy forecasting.

We here focus on solar energy which is one of the most important sources of renewable energy. For example, photovoltaic (PV) power contributes significantly to the power supply

in Germany and generated 8.2% of the gross electricity consumption in 2019, and temporarily up to 50% of the current electricity consumption on sunny days (Fraunhofer Institute for Solar Energy Systems, 2020). Solar energy forecasting approaches can be distinguished into those that aim to predict solar irradiance, and those that aim to predict PV power. Naturally, solar irradiance and PV system output are strongly correlated, and the employed statistical methods are similar (van der Meer et al., 2018). We will focus on probabilistic solar irradiance forecasting in the following.

For recent comprehensive overviews and reviews of existing approaches, see van der Meer et al. (2018) and Yang (2019). Except for short-term prediction (e.g., Zelikman et al., 2020), most methods for probabilistic solar irradiance forecasting combine physical information from NWP models with statistical methods.

5.2.1 Data

The postprocessing models are applied to two datasets that focus on distinct solar irradiance variables, NWP models, temporal resolutions and geographic regions (Hungary and Germany), for lead times of up to 48 and 120 hours, respectively.

AROME-EPS

The 11-member *Applications of Research to Operations at Mesoscale EPS* (AROME-EPS) of the *Hungarian Meteorological Service* (HMS) covers the Transcarpatian Basin with a horizontal resolution of 2.5 km (Jávorné Radnóczi et al., 2020). It consists of 10 ensemble members obtained from perturbed initial conditions and a control member from an unperturbed analysis. The dataset at hand contains ensemble forecasts of instantaneous values of *global horizontal irradiance* (GHI) (W/m^2) together with the corresponding validation observations of the HMS for seven representative locations in Hungary (Aszód, Budapest, Debrecen, Kecskemét, Pécs, Szeged, Tápiószele) for the period between 7 May 2020 and 14 October 2020. Forecasts are initialized at 00 UTC with a forecast horizon of 48 hours and a temporal resolution of 30 minutes resulting in a total of 96 forecasts per submission. For the AROME-EPS, we will refer to the term lead time as the time between the initialization and the time stamp of the corresponding instantaneous forecast.

ICON-EPS

A general introduction of the ICON data that is used for postprocessing on the KIT-Weather portal is provided in Section 5.1.1. For the solar irradiance variables, we use forecasts that are initialized four times a day at 00, 06, 12 and 18 UTC each with a forecast horizon of 120 hours.[15] The ICON ensemble predictions are given as averages between two time stamps, e.g., the 12-step ahead forecast is the average predicted irradiance between 11 to 12 hours after

[15]As for the mean sea level pressure, no ICON-Global forecasts are available for the solar irradiance variables.

initialization time and the 59-step ahead forecast is the average from 84 to 90 hours. For simplicity, we will refer to an individual forecast by the lead time and not the step ahead, where lead time refers to the time between submission and the final time stamp, i.e., the former forecast has a lead time of 12 hours, the latter a lead time of 90 hours.

Our dataset contains ensemble forecasts of the two components of GHI: *beam normal irradiance* (BNI) adjusted for the solar zenith angle θ (i.e., $\text{BNI} \cdot \cos(\theta)$), and *diffuse horizontal irradiance* (DHI) (W/m^2). To simplify the distinction between the different types of irradiance and improve the readability of this section, we will refer to $\text{BNI} \cdot \cos(\theta)$ as *beam horizontal irradiance* (BHI) or direct irradiance, to DHI as diffuse irradiance, and to $\text{GHI} = \text{BHI} + \text{DHI}$ as global irradiance, in the following.

As described in Section 5.2.1, we further obtained corresponding observational data for weather stations located near the major cities of BE, HH and KA. The observations are computed based on 10-minute sums of the corresponding variables. For detailed descriptions of the observations, we refer again to Becker and Behrens (2012). In contrast to Section 5.1, the entire dataset used here covers the period from 27 December 2018 to 31 December 2020, which refers to the data available, when the study presented in this section was conducted.

5.2.2 Model configurations

In this section, we first highlight the choice of forecast distribution before specifying the EMOS models used and how we utilize the underlying datasets. At last, we describe the configurations used for each of the two datasets.

Choice of forecast distribution

As indicated in Section 4.2.1, the discrete-continuous nature of solar irradiance calls for nonnegative predictive distributions assigning positive mass to the event of zero irradiance. Similar to parametric approaches to postprocessing ensemble forecasts of precipitation accumulation, one can either left-censor an appropriate continuous distribution at zero (e.g., Scheuerer, 2014; Baran and Nemoda, 2016), or choose the more complex method of mixing a point mass at zero and a suitable continuous distribution with nonnegative support (Sloughter et al., 2007; Bentzien and Friederichs, 2012). Here, focus on the former and use a zero-censored logistic distribution (CL0) in line with the precipitation models of Section 5.1.

Note that initial tests with a censored normal predictive distribution were performed for the ICON-EPS dataset; however, the results suggested that the proposed CL0-EMOS approach results in slightly improved predictive performance. The choice of parametric families for the forecast distribution has been an important aspect in postprocessing research. For considerations in the context of solar irradiance forecasting, see e.g., Yang (2020b), Yagli et al. (2020), and Le Gal La Salle et al. (2020). In those previous works, forecast distributions are truncated at zero.

Our choice of a left-censored distribution was motivated by the aim to obtain a single distributional model that can be applied to all times of day and is able to account for cases where the observation and all or most of the ensemble member predictions are zero, which makes it unnecessary to remove nighttime irradiance data during training and inference. Therefore, there is no need to select location- and season-specific times of day that define periods of time where the postprocessing model can be applied, which is the case for models based on truncated distributions. In addition, we have observed occasional cases where zero irradiance is observed, but some of the ensemble members predict nonzero values. A model based on a censored distribution is able to correct those deficiencies and might have advantages for applications such as the KIT-Weather portal or automated procedures where postprocessing constitutes one of the components and postprocessed forecasts serve as inputs for additional modeling steps.

EMOS models for solar irradiance forecasting

On the KIT-Weather portal, we are using the same EMOS model as for the precipitation variables to postprocess the solar irradiance variables (Table 5.2). Based on the KIT-Weather approach, we derive the models used in this case study starting with a simple EMOS model, where the location parameter μ and the scale parameter σ of the CL0-distribution are connected to the ensemble members via link functions

$$\mu = a + b_0 p_0(x) + b_1 x_1 + \cdots + b_m x_m \qquad \text{and} \qquad \sigma = \exp\Big(c + d \log s^2(x)\Big), \tag{5.3}$$

where the EMOS coefficients $a, b_0, b_1, \ldots, b_m$ and c, d are estimated according to the optimum score principle. Note that we use the fraction of zero-ensemble members as additional predictor variable to the ensemble members analogous to the parameterization of the precipitation models in Table 5.2.

In order to capture the seasonal variation in solar irradiance, following the ideas of Hemri et al. (2014), we further fit separate periodic models to both observations and ensemble forecasts of the training data. Two regression models dealing with oscillations of a single and two different frequencies are investigated, namely

$$y_t = \alpha_0 + \alpha_1 \sin\left(\frac{2\pi t}{365}\right) + \alpha_2 \cos\left(\frac{2\pi t}{365}\right) + \varepsilon_t \qquad \text{and} \tag{5.4}$$

$$y_t = \beta_0 + \beta_1 \sin\left(\frac{2\pi t}{365}\right) + \beta_2 \cos\left(\frac{2\pi t}{365}\right) + \beta_3 \sin\left(\frac{4\pi t}{365}\right) + \beta_4 \cos\left(\frac{4\pi t}{365}\right) + \varepsilon_t, \tag{5.5}$$

where the dependent variables y_t, $t = 1, 2, \ldots, n$, are either irradiance observations for a given location or members of the corresponding ensemble forecast with a given lead time h from a training period of length n. With the help of either equation (5.4) or (5.5) one can calculate the h ahead predictions $\widehat{y}$ and $\widehat{x}_i$ of the observation and ensemble members, respectively, and

consider the following modified link function for the location:

$$\mu = \widehat{y} + a + b_0 p_0(x) + b_1 \left(x_1 - \widehat{x}_1\right) + \cdots + b_m \left(x_m - \widehat{x}_m\right). \tag{5.6}$$

The model formulations in equations (5.3) and (5.6) are valid under the assumption that each ensemble member can be identified and tracked. However, most operationally used EPSs today generate ensemble forecasts that lack individually distinguishable physical features such as distinct variations in the model physics, for example by generating ensemble member based on random perturbations of initial conditions. Those statistically indistinguishable members (or groups of members) generated in this way are usually referred to as exchangeable (Fraley et al., 2010) in reference to the concept of exchangeable random variables in statistics. This is also the case for the ICON-EPS and the AROME-EPS described in Section 5.2.1. The existence of groups of exchangeable ensemble members should be taken into account during model formulation. This is usually achieved by requiring that ensemble members within a given group share the same coefficients (e.g., Wilks, 2018). If there exist m ensemble members divided into K exchangeable groups and $\overline{x}_k$ denotes the mean of the kth group containing m_k ensemble members ($\sum_{k=1}^{K} m_k = m$), the exchangeable versions of the link functions in equations (5.3) and (5.6) are

$$\mu = a + b_0 p_0(x) + b_1 \overline{x}_1 + \cdots + b_K \overline{x}_K \tag{5.7}$$

and

$$\mu = \widehat{y} + a + b_0 p_0(x) + b_1 \left(\overline{x}_1 - \widetilde{x}_1\right) + \cdots + b_K \left(\overline{x}_K - \widetilde{x}_K\right), \tag{5.8}$$

respectively, where $\widetilde{x}_k$ is the prediction of $\overline{x}_k$ for lead time h based either on equation (5.4) or (5.5) for $k = 1, \ldots, K$.

Training data selection

In the following case study, examples of all training data selection methods listed in Section 5.1.1 are shown: global estimation with a rolling training period for the AROME-EPS and local estimation with rolling and extending training periods for the ICON-EPS. Note that this includes the selection used on the KIT-Weather portal for the ICON data, that is, local estimation with an extending training period. Further, we follow the training data selection described in Section 5.1.1 by treating each lead time and each initialization hour separately. Since we do not remove nighttime data during training and inference, this further helps to account for positive probabilities of observing zero irradiance as point masses in the forecast distributions. Note that seasonal variations for a given time of day, for example effects of differing solar zenith angles, are implicitly modeled when using equation (5.4) or (5.5).

Configuration for AROME-EPS

As discussed in Section 5.2.1, the AROME-EPS consists of a control member and 10 exchangeable ensemble members obtained using perturbed initial conditions. The dataset at hand covers a short time period only, in particular compared to the ICON-EPS dataset, with forecast-observation pairs available for only 159 calendar days. Therefore, the available training periods cannot be long enough for accurate modeling of seasonal oscillations and we only consider a CL0-EMOS model where the location is linked to the ensemble members via equation (5.7) with $K = 2$ and $m_1 = 1$, $m_2 = 10$, which means that six parameters need to be estimated. We consider global estimation with a rolling training period of length 31 days, leaving 127 calendar days (9 June 2020 to 13 October 2020) for forecast verification, and refer to this model as the *simple RT* model. The choice of the training period length corresponds to typical values in the postprocessing literature and was made to have a similar forecast case per parameter ratio as for the best performing model for the ICON-EPS data. In light of the limited size of the dataset, it is not surprising that the use of monthly expanding training periods or local parameter estimation procedures results in worse predictive performance, and we omit the corresponding results in the interest of brevity.

Configuration for ICON-EPS

In contrast to the AROME-EPS, the ICON-EPS dataset covers a substantially longer time period and therefore allows for considering and comparing more complex model formulations and estimation procedures. As members of the ICON-EPS are obtained with the help of random perturbations, they can be regarded as exchangeable. Hence, for postprocessing we use the CL0-EMOS model with locations linked to the ensemble members either via equation (5.7) or (5.8) with $K = 1$. Thus, for the model of equation (5.3) with the location parameter of equation (5.7) (which was the only model variant considered for the AROME-EPS and is referred to as *simple model*) one has to estimate five unknown parameters, whereas more complex approaches, which account for seasonal variations in the link function in equation (5.8) of the location parameter via equation (5.4) (referred to as *periodic model*) or (5.5) (referred to as *periodic 2 model*), require the estimation of a total of 11 and 15 parameters, respectively.

The period from 27 December 2018 to 31 December 2019 is used for training purposes only, the calendar year 2020 (366 calendar days) for model verification, which leaves enough flexibility for choosing a sufficiently long training period even for local modeling. Two different training configurations are investigated: a rolling training (RT) period of length 365 days, and a monthly expanding training (MET) scheme, where all data until the end of the last month before the forecast date under consideration is used for training. In the latter case, the first training period includes all data prior to calendar year 2020. According to initial studies (not shown), MET provides reasonable verification scores only for the simple model. Therefore, we

report results for the simple model with rolling (*simple RT*) and monthly expanding training (*simple MET*) as well as for the periodic models with rolling training (*periodic RT* and *periodic 2 RT*).

5.2.3 RESULTS

In the following case studies, the forecast skill of various variants of the CL0-EMOS model is evaluated. First, we consider a simple CL0-EMOS variant for the HMS AROME-EPS ensemble forecasts of GHI, then we will investigate the performance of the more complex models for the DWD ICON-EPS ensemble forecasts of direct and diffuse irradiance.

AROME-EPS

Here, we will asses the predictive performance of the simple RT model described above. Recall that the ensemble predictions of GHI are provided at a temporal resolution of 30 minutes. As all AROME-EPS forecasts are initialized at 00 UTC, the forecast lead time either coincides with the time of observation or has a shift of 24 hours. Hence, all scores are reported as functions of the lead time. We further average over results from all seven observation locations over Hungary.

Note that in contrast to considering predictions of GHI directly, the standard approach in solar forecasting is the use of a *clear-sky index* (CSI) as target variable to stationarize the time series of irradiances (van der Meer et al., 2018; Yang, 2020a; Yang et al., 2020). The clear-sky irradiance used for the normalization is obtained from clear-sky models which estimate the amount of solar radiation arriving at the surface under clear-sky (cloud-free) conditions, see Yang (2020a) for an in-depth discussion and comparison of available models. To investigate the differences between postprocessing forecasts of GHI and forecasts of CSI, we follow the procedure outlined in Yang (2020b, Section II.A) to convert the GHI ensemble predictions $\{x_1, x_2, \ldots, x_m\}$ and the GHI observation y to CSI values. To do so, we obtained clear-sky irradiance values from the McClear model using the `camsRad` package (Lundstrom, 2016) for the locations and relevant time instances, and converted GHI to CSI by division by the corresponding clear-sky irradiance. We then used an identical model formulation and training procedure as for GHI, and derived 100 equidistant quantiles from the postprocessed forecast distributions for CSI. Those quantiles were transformed back to GHI values by multiplying with the corresponding clear-sky irradiance and used for approximating the verification scores. We refer to this approach as *simple RT CSI*.

Figure 5.10 shows the mean CRPS of calibrated and raw ensemble forecasts and the CRPSS with respect to the raw ensemble. Postprocessing using the simple RT approach improves the forecast performance when positive irradiance is likely to be observed (03–19 UTC), and performs no corrections otherwise, resulting in a skill score of zero. Note that compared with direct calibration of the GHI, postprocessing of the CSI predictions does not result in a

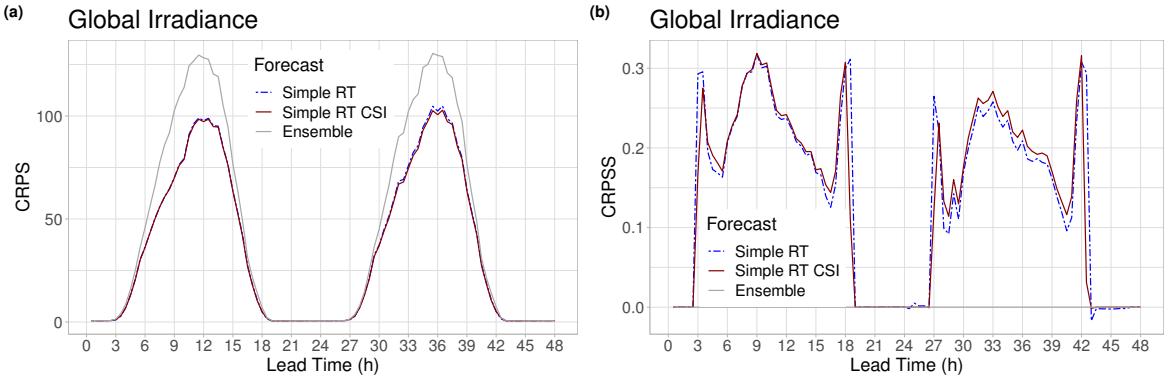

Figure 5.10: Mean CRPS of postprocessed and raw ensemble forecasts of GHI (a) and CRPSS with respect to the raw ensemble (b) as functions of lead time for the AROME-EPS dataset.

substantial difference or clear improvement in forecast skill. These observations are in line with the results reported in Yang (2020b) in a related context. Hence, in the remainder only the results for the former approach will be reported. To assess the statistical significance of the improvements in predictive performance compared to the raw ensemble predictions, we performed a block bootstrap resampling to compute 95% confidence intervals (Figure 5.11). The corresponding standard deviations are obtained from 2,000 block bootstrap samples calculated using the stationary bootstrap scheme, where the mean block length is computed according to Politis and Romano (1994). We found that the observed improvements are statistically significant, with skill scores of postprocessed forecasts being significantly positive for all time periods where positive irradiance is likely to be observed. The large jumps in the CRPSS at 4, 19, 27 and 42 hours are mainly caused by numerical issues as at these lead times the mean CRPS of both raw and postprocessed forecasts is very close to 0, and also leads to an increased width of the confidence intervals in Figure 5.11. For qualitatively similar observations in a related context, see Bakker et al. (2019, Figure 7).

To this point, the relative improvements of the postprocessed forecasts in terms of the CRPSS were investigated by computing the corresponding skill score values of the CRPSS using the ensemble forecast as reference (Figures 5.10). To further investigate the relative improvements in comparison to a more naive reference, Figure 5.11 shows corresponding values of the CRPSS using a climatological forecast as reference model, where observations of the rolling training period are considered as a forecast ensemble. This can be viewed as a persistence ensemble in the terminology of Yang (2019). Here, the individual climatological forecasts are based on observations of the preceding 31 days. Both ensemble and the postprocessed forecasts show clear improvements for times of day during which it is unlikely to observe zero irradiance. Corresponding skill scores for MAE and RMSE indicate a very similar behavior of the deterministic forecasts and are not shown here in the interest of brevity.

A similar behavior can be observed for the *Brier skill score* (BSS) values of threshold

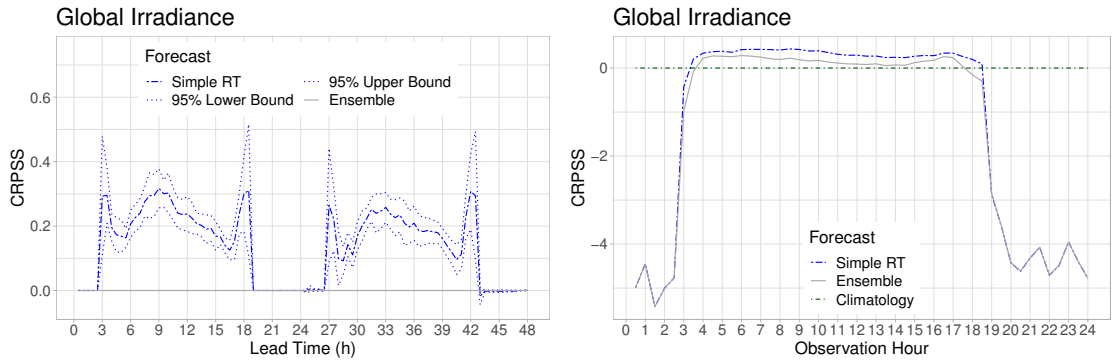

Figure 5.11: CRPSS of EMOS postprocessed forecasts with respect to the raw ensemble together with 95% confidence intervals (left), and CRPSS of postprocessed and raw ensemble forecasts with respect to climatology as functions of the observation hour (right), for the AROME-EPS dataset.

exceedance shown in Figure 5.12, where the threshold values correspond to the 40th, 60th, 90th and 95th percentiles of observed nonzero GHI (25, 127, 498, 604 $\mathrm{W/m^2}$). The results are consistent in that the higher the threshold, the shorter the period with a positive mean BS, as the higher thresholds are mostly observed around midday, when the irradiance is strongest. For the corresponding lead times, the postprocessed forecasts outperform the raw ensemble. Again, negative skill scores appear only at the boundaries where the mean score values to be compared are very small.

Figure 5.13a showing the coverage of 83.33% PIs further confirms the improved calibration of the postprocessed forecast. Between 03 and 19 UTC, when positive GHI is likely to be observed, the EMOS model results in a coverage close to the nominal value, whereas the coverage of the raw ensemble is consistently below 60%.

Further, Figure 5.13b showing the MAE of the median forecasts indicates that postprocessing substantially improves the accuracy of point forecasts as well. At the hours of peak irradiance the difference in MAE exceeds 20 $\mathrm{W/m^2}$. As we will see below, this is in a strong contrast with the results of the second case study (Figure 5.23) and indicates the presence of a bias in the AROME-EPS that is alleviated by postprocessing. Similar conclusions can be drawn from the RMSE of the mean forecasts (not shown).

The presence of a bias in the raw ensemble forecasts can also be observed in the verification rank histograms shown in Figure 5.14. In addition, the clearly U-shaped verification rank histograms indicate a strong underdispersion, which is in line with the low coverage of the raw ensemble forecasts observed in Figure 5.13a and persists across all considered ranges of lead times. However, the ensemble members are more likely to underestimate the true irradiance, which further indicates a negative bias. Both deficiencies are successfully corrected by statistical postprocessing. The PIT histograms of EMOS predictive distributions given in the upper row of Figure 5.14 are almost flat indicating just a minor bias for observations between 12:30 and 24 UTC.

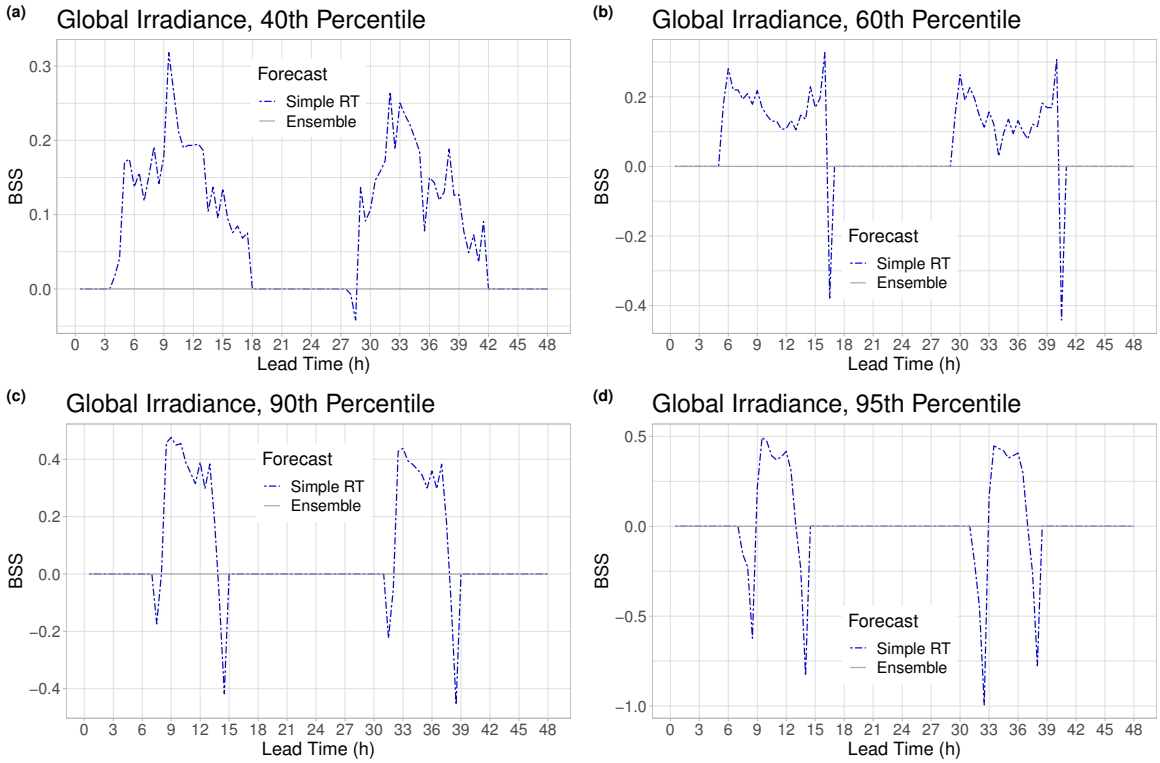

Figure 5.12: BSS of postprocessed forecasts with respect to the raw ensemble as function of lead time for the AROME-EPS dataset, with thresholds corresponding to the 40th, 60th, 90th and 95th percentiles of observed nonzero GHI.

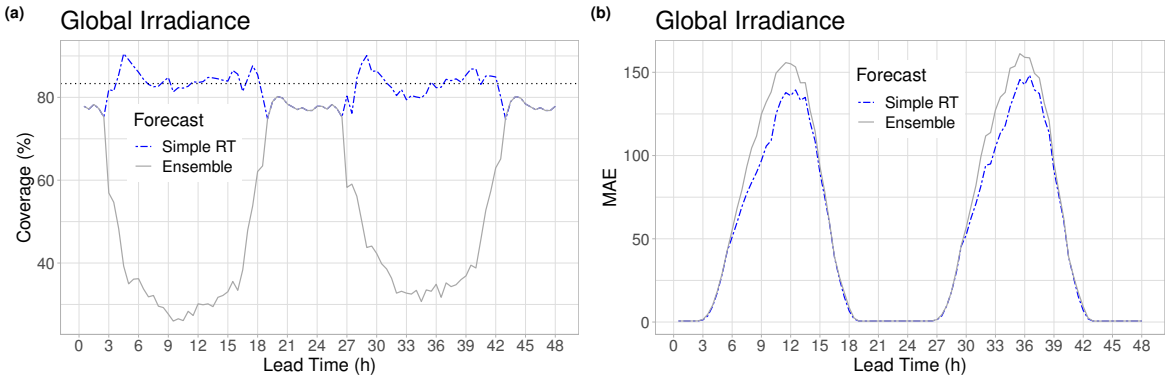

Figure 5.13: Coverage of 83.33% PIs of postprocessed and raw forecasts (a), and MAE of the median forecasts (b) for the AROME-EPS dataset.

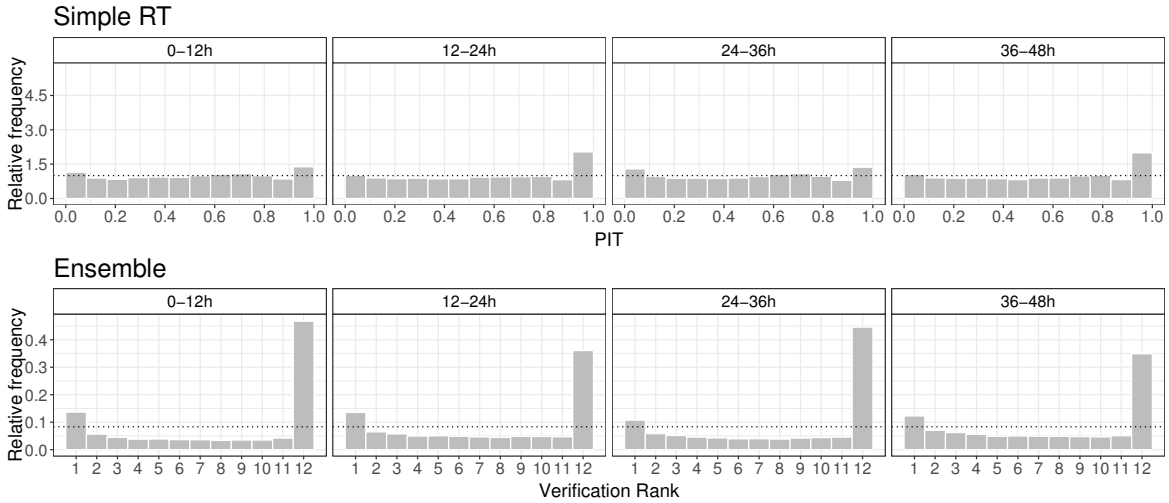

Figure 5.14: PIT histograms of postprocessed and verification rank histograms of raw ensemble forecasts of GHI for the lead times 0–12, 12–24, 24–36 and 36–48 hours.

ICON-EPS

Recall that the ICON-EPS dataset covers a substantially longer time period than the AROME-EPS and therefore we consider and compare more complex model formulations and estimation procedures, namely, the simple MET, the simple RT, the periodic RT and the periodic 2 RT model. Further, we here consider forecasts of direct irradiance (BHI) and diffuse irradiance (DHI) at temporal resolutions of 1 hour (for lead times up to 48 hours), 3 hours (for lead times 51–72 hours) and 6 hours (for lead times 78–120 hours), resulting in a forecast horizon of 120 hours. Given the negligible differences we observed when comparing postprocessing of GHI and CSI for the AROME-EPS data, we only consider predictions of BHI and DHI without normalization by the corresponding clear-sky irradiances here.

Raw ensemble forecasts of direct and diffuse irradiance are used as references models. Unless indicated otherwise, results discussed below are averaged over all three observation locations and all four initialization times of the NWP model. Note that this might make the interpretation of the results more involved than in the first case study due to the interacting effects of forecast initialization time, lead time, and corresponding time of day of the observation.

First, we investigate diurnal effects by examining the dependence of the mean CRPS of the various forecast models on the time of the observation shown in Figure 5.15a,b. In order to provide a fair comparison, we take only the first 48 hours of the forecast horizon into account, where hourly forecasts are available. Both for BHI and DHI, all postprocessing methods outperform the raw ensemble forecasts at all time points when positive irradiance is likely to be observed. According to the skill scores with respect to the raw ensemble shown in Figure 5.15c, in the case of direct irradiance, the predictive performance mainly depends on the complexity of model formulations and parameter estimation, with more complex models

exhibiting better forecast performance. However, the differences between the various EMOS approaches are relatively minor. The same applies for diffuse irradiance in early and late hours (Figure 5.15d), whereas between 06–18 UTC there is no visible difference in the skill of the different EMOS models.

Analogous to the AROME-EPS, we also compare against a climatological reference model (Figure 5.16). Whereas the climatology is based on observations of the preceding 31 days in case of the AROME-EPS, the ensemble has 365 members for the ICON-EPS. Again, both ensemble and postprocessed forecasts clearly improve the reference when zero irradiance is unlikely to observe, and we skip corresponding skill scores for MAE and RMSE that indicate a similar behavior in the interest of brevity.

Note that the apparent periodic oscillations in the CRPSS values might be partly caused by the pooling of different observation hours due to the four considered initialization times. In contrast to the AROME-EPS, postprocessing also improves the predictive performance at night achieving a CRPS of almost zero. ICON-EPS fails to achieve mean CRPS values of zero due to occasional predictions of nonzero irradiance values during night times.

Figure 5.17a, which shows the CRPSS with respect to the raw BHI ensemble forecasts as function of the lead time, confirms the observations from Figure 5.15c. Here, the differences are more pronounced due to the different scaling of the vertical axis, and again, the periodic 2 model with rolling training period exhibits the best forecast skill, whereas the simple model with monthly expanding training shows the smallest CRPSS. In general, all skill scores decrease for longer lead times, which also holds for the corresponding CRPSS values for DHI (Figure 5.17b). Overall, slightly larger improvements relative to the raw ensemble are observed for direct than for diffuse irradiance, and none of the models result in negative skill scores. Up to a lead time of 48 hours, there are no visible differences between the various EMOS approaches. For longer lead times, similar to BHI, the most complex periodic 2 model shows the best predictive performance, whereas the simple model with parameters estimated using a rolling training period is now the least skillful. Recall that for longer lead times the forecasts refer to a longer time period and thus seasonal effects regarding the diurnal cycle might be captured by the more complex models.

Analogously to Figure 5.11, we assess the statistical significance of the improvements of the best performing periodic 2 RT EMOS model compared to the raw ICON-EPS forecasts using the corresponding 95% confidence intervals. In the case of BHI, postprocessing results in a significant improvement in mean CRPS up to 60 hours ahead, whereas postprocessed forecasts of DHI significantly outperform the raw ensemble over the entire forecast horizon of 120 hours. Further, comparing Figures 5.17 and 5.18 it can be observed (especially in the case of diffuse irradiance) that there is no significant difference between the various postprocessing methods in terms of mean CRPS.

A third aspect is the dependence of the forecast skill on the observation location. Table 5.5 shows the overall CRPSS of the different EMOS models with respect to the raw forecasts and

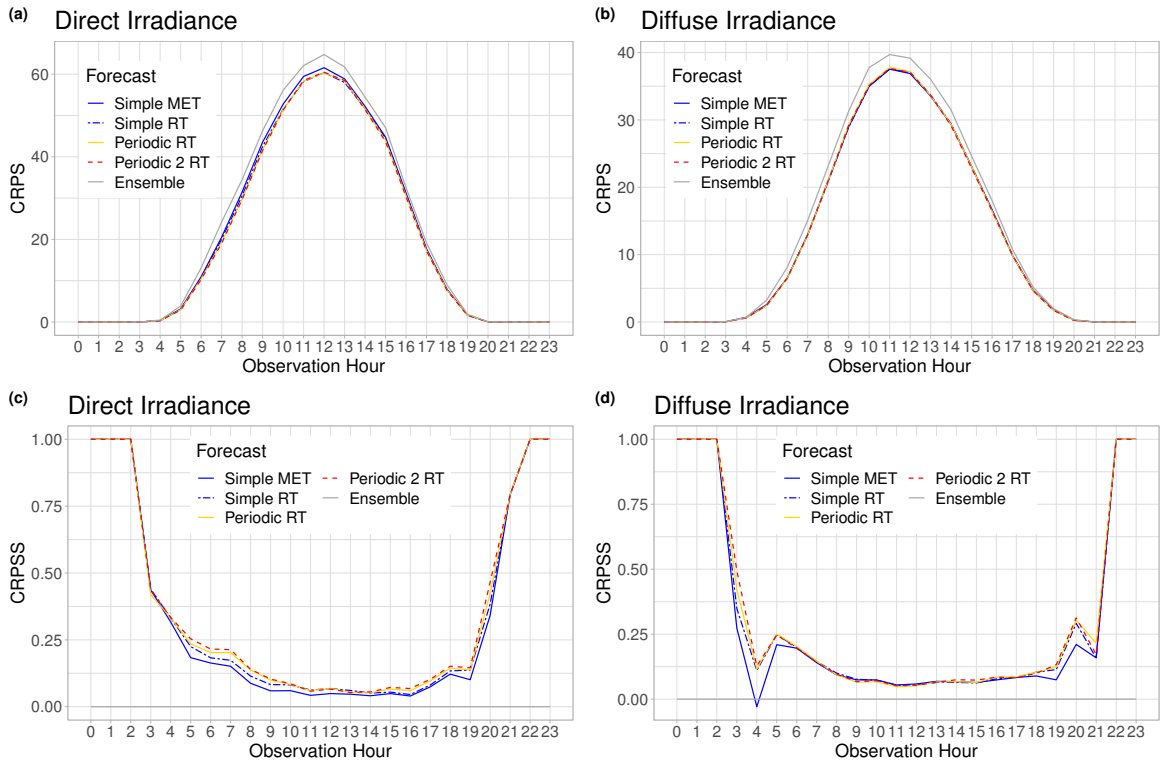

Figure 5.15: Mean CRPS of postprocessed and raw ensemble forecasts of direct (a) and diffuse (b) irradiance, and corresponding skill scores (c,d) with respect to the raw ensemble as functions of the observation hour for the ICON-EPS dataset.

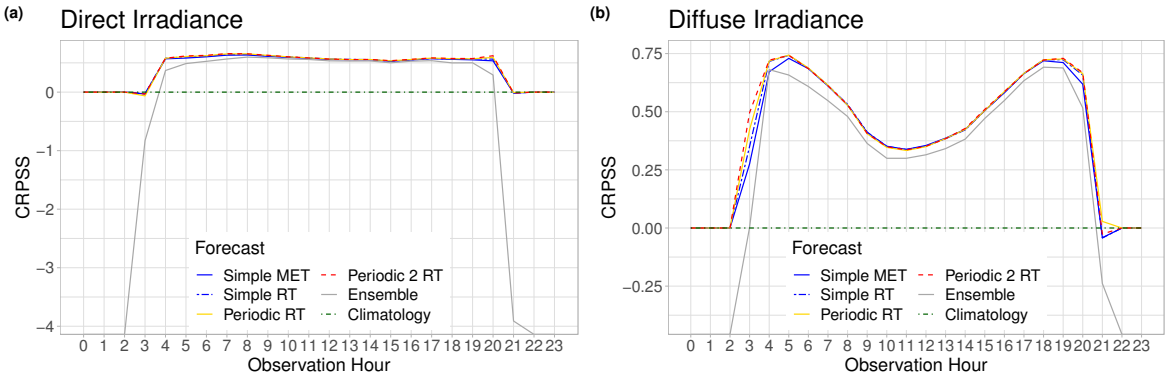

Figure 5.16: CRPSS of postprocessed and raw ensemble forecasts of direct (a) and (b) diffuse irradiance with respect to climatology as functions of the observation hour for the ICON-EPS dataset.

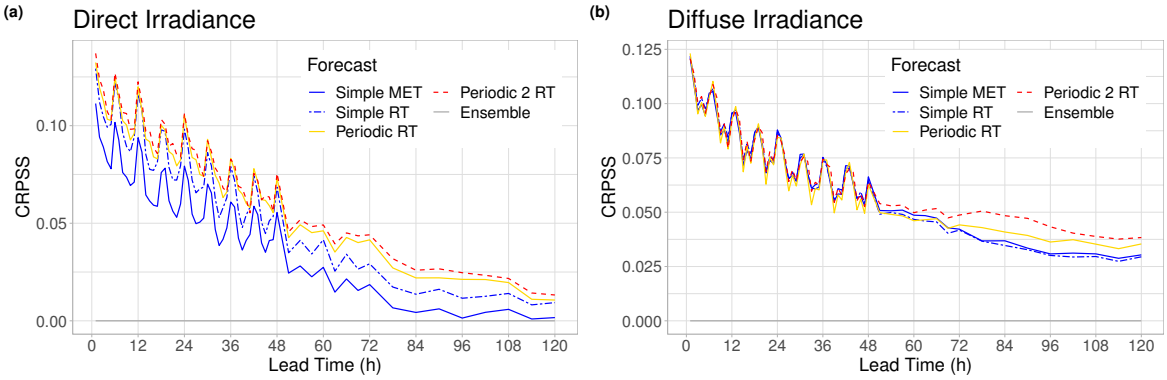

Figure 5.17: CRPSS of postprocessed forecasts of direct (a) and diffuse (b) irradiance with respect to the raw ensemble as functions of the lead time for the ICON-EPS dataset.

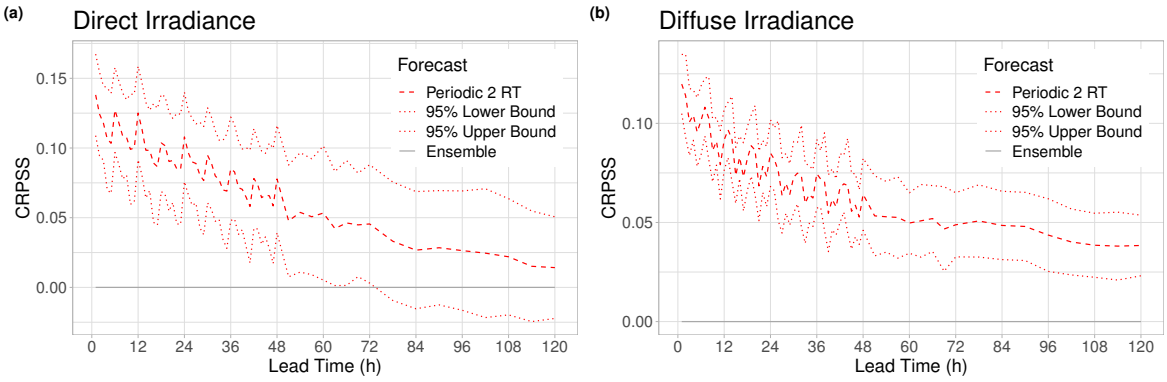

Figure 5.18: CRPSS of the best performing postprocessed forecasts of direct (a) and (b) diffuse irradiance with respect to the raw ensemble together with 95% confidence intervals for the ICON-EPS dataset.

the corresponding CRPSS values of the three different cities for four different intervals of the forecast horizon. The main message of these results is that the magnitude of improvements in predictive performance resulting from postprocessing strongly depends on the location. For both variables, KA benefits the most, while for BE after 24 hours, and for HH after 78 hours the simple MET model performs worse than the raw BHI ensemble forecast and results in negative skill scores. Among the competing models for BHI the most complex periodic 2 RT model shows the best forecast skill for BE and HH, and shows the best overall performance as well. In the case of DHI, the differences in performance between the various EMOS models are much smaller, which is in line with the results observed in Figure 5.17b. In particular, none of the more complex models consistently outperforms the simple MET model.

To investigate seasonal effects in the improvements achieved via postprocessing, Figure 5.19 shows the CRPSS of the postprocessed forecasts based on monthly mean values. For direct irradiance, the improvements are generally larger in winter than in summer. From November

Lead Time		Direct Irradiance				Diffuse Irradiance			
in hours	Model	Overall	KA	BE	HH	Overall	KA	BE	HH
	Simple MET	0.076	0.114	0.029	0.082	0.089	0.099	0.075	0.093
1–24	Simple RT	0.095	0.130	0.070	0.083	0.091	0.100	0.080	0.092
	Periodic RT	0.101	0.133	0.077	0.090	0.089	0.099	0.080	0.086
	Periodic 2 RT	0.104	0.132	0.082	0.097	0.091	0.097	0.083	0.092
	Simple MET	0.050	0.091	-0.000	0.054	0.066	0.076	0.050	0.070
25–48	Simple RT	0.064	0.102	0.032	0.054	0.066	0.075	0.053	0.069
	Periodic RT	0.072	0.102	0.045	0.066	0.064	0.071	0.055	0.063
	Periodic 2 RT	0.074	0.099	0.047	0.073	0.066	0.069	0.060	0.069
	Simple MET	0.021	0.062	-0.018	0.018	0.048	0.055	0.041	0.046
51–72	Simple RT	0.033	0.071	0.008	0.020	0.046	0.053	0.041	0.043
	Periodic RT	0.043	0.075	0.021	0.030	0.047	0.053	0.045	0.041
	Periodic 2 RT	0.046	0.069	0.024	0.043	0.051	0.049	0.055	0.049
	Simple MET	0.004	0.028	-0.016	-0.002	0.032	0.032	0.039	0.025
78–120	Simple RT	0.013	0.036	-0.001	0.001	0.031	0.031	0.040	0.021
	Periodic RT	0.019	0.032	0.017	0.008	0.038	0.037	0.048	0.026
	Periodic 2 RT	0.022	0.024	0.021	0.022	0.043	0.036	0.062	0.030

Table 5.5: Overall CRPSS and CRPSS for individual locations of postprocessed forecasts of direct and diffuse irradiance with respect to the raw ensemble.

to April, the differences among the postprocessing approaches are most pronounced, and more complex model formulations that incorporate seasonal effects particularly show improved performance. For diffuse irradiance, the overall level of improvements in terms of the mean CRPS is smaller (note the different scale of the vertical axes). Only minor seasonal effects in the form of smaller improvements between October and December can be detected.

To simplify the presentation of the results, in the remaining part of this section we consider pooled data of all locations, months and observation hours and display the dependence on the lead time only. The improved calibration of postprocessed forecasts can also be observed in the coverage plots of Figure 5.20. All postprocessing approaches result in a coverage close to the nominal 95.12% for all lead times, whereas the maximal coverage of the raw ensemble is below 85% for both variables. The difference between postprocessed forecasts of BHI is more pronounced with periodic models being the closest to the nominal value. These results are in line with the shapes of the PIT and verification rank histograms of Figures 5.21 and 5.22. Raw ensemble forecasts of BHI are strongly underdispersed and slightly biased for all lead times. Even though this is slightly alleviated for longer lead times, the PIT histograms of all EMOS models are much closer to the desired uniform distribution. However, some bias still remains in the postprocessed forecasts. In contrast, neither the PIT histograms of postprocessed, nor the verification rank histograms of raw forecasts of DHI indicate any bias

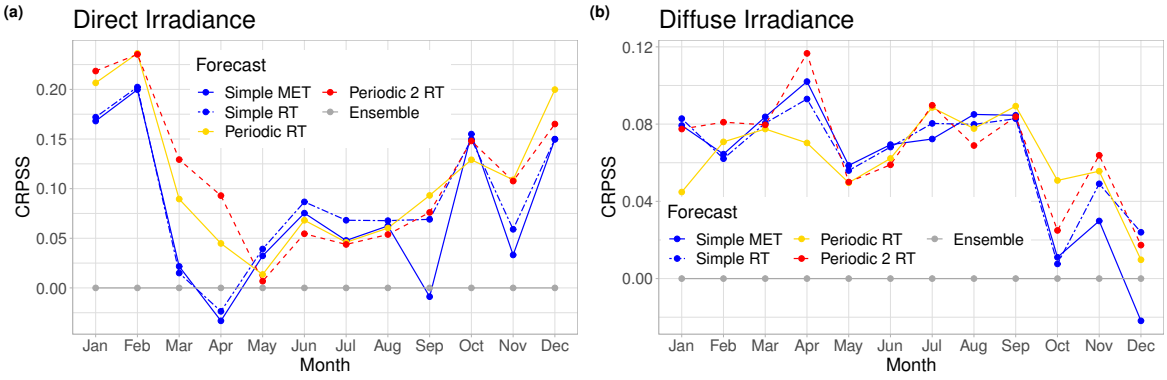

Figure 5.19: CRPSS of postprocessed forecasts of direct (a) and diffuse (b) irradiance with respect to the raw ensemble, computed based on monthly mean values for the ICON-EPS dataset.

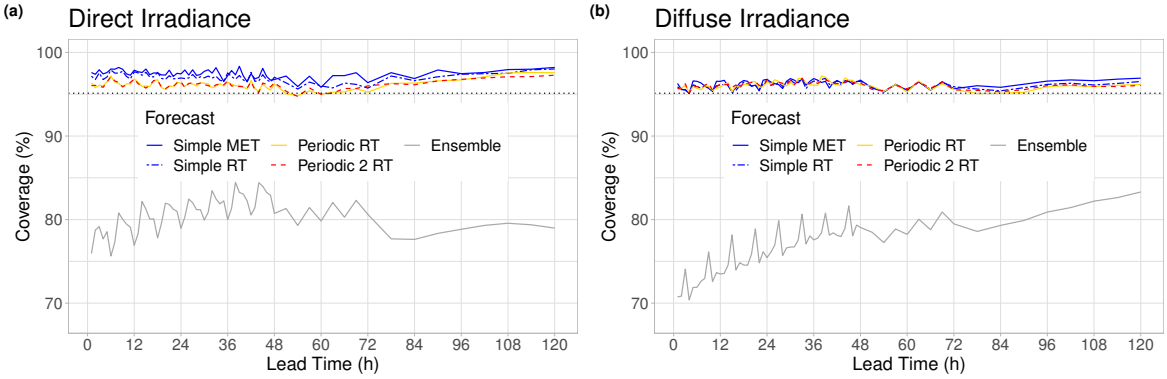

Figure 5.20: Coverage of 95.12% PIs of postprocessed and raw forecasts of direct (a) and diffuse (b) irradiance for the ICON-EPS dataset.

(Figure 5.22), and all EMOS approaches successfully correct the underdispersion of the raw ensemble resulting in almost perfectly uniform PIT histograms.

Finally, Figure 5.23 showing the MAE of the median forecasts indicates that while postprocessing substantially improves the calibration of probabilistic forecasts, it has a minor effect on the accuracy of point forecasts. The difference in MAE is less than $2\,\mathrm{W/m^2}$ for direct irradiance and $0.6\,\mathrm{W/m^2}$ for diffuse irradiance for all considered lead times. The sharp changes in MAE values at 51 and 78 hours are results of the change in temporal resolution of the forecasts. Corresponding results for the RMSE of the mean forecasts are very similar and thus not shown here.

Conclusions

We propose a postprocessing method for ensemble forecasts of solar irradiance based on the EMOS approach. Several model variants that differ in terms of the temporal composition of training datasets and adjustments to seasonal variations in the model formulation are

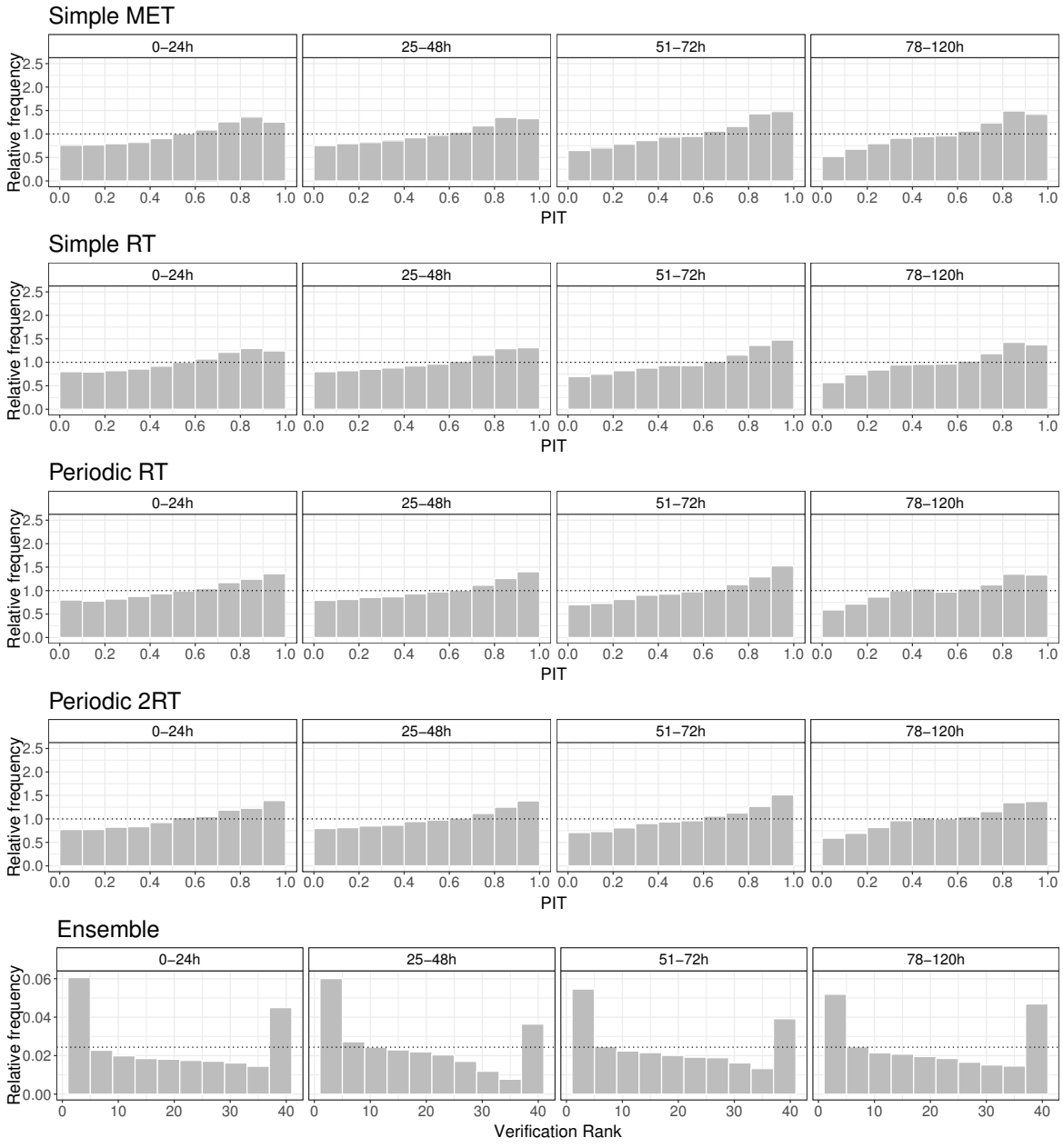

Figure 5.21: PIT histograms of postprocessed and verification rank histograms of raw ensemble forecasts of DNI for the lead times 0–24, 25–48, 51–72 and 78–120 hours for the ICON-EPS dataset.

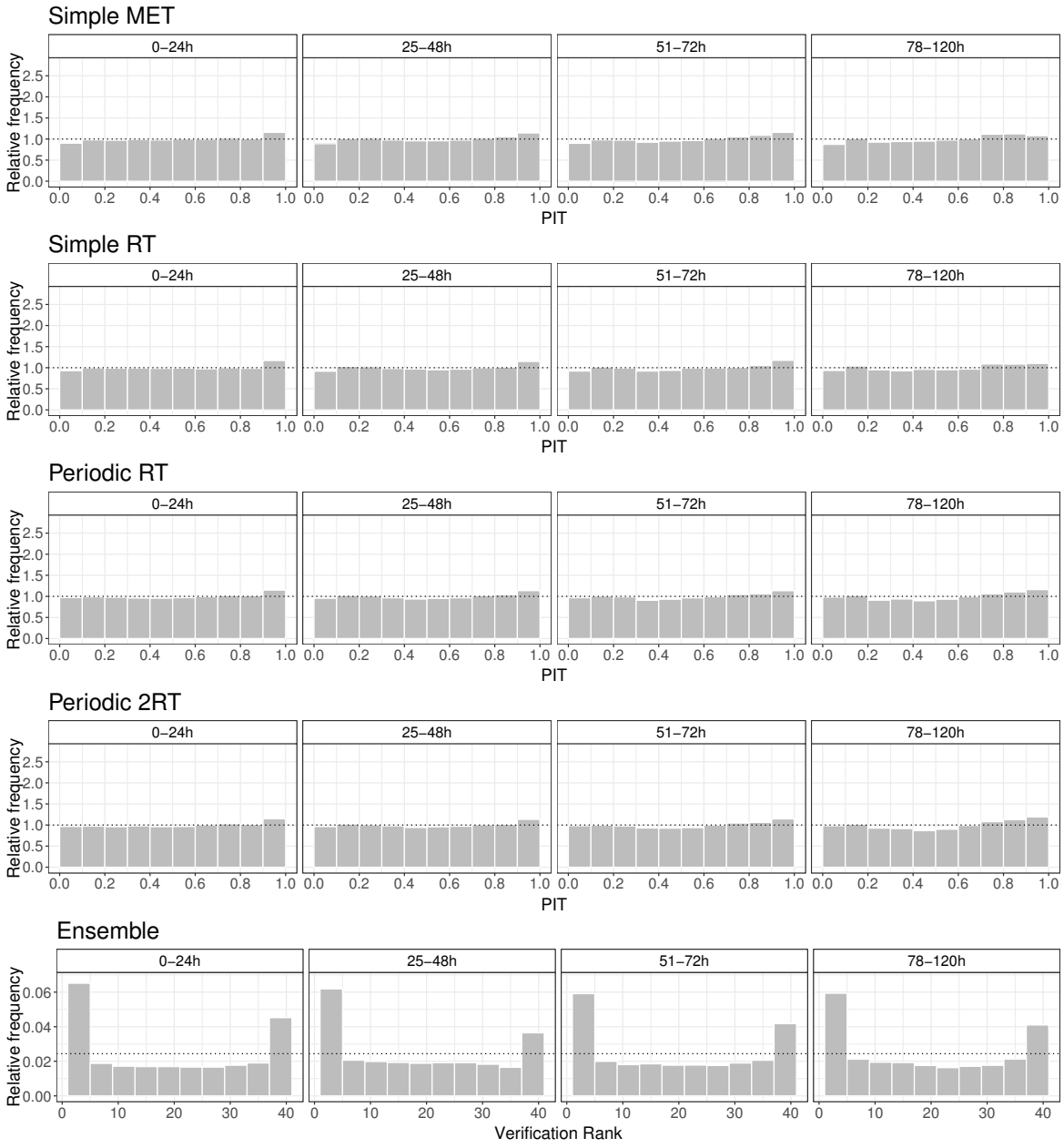

Figure 5.22: PIT histograms of postprocessed and verification rank histograms of raw ensemble forecasts of DHI for the lead times 0–24, 25–48, 51–72 and 78–120 hours for the ICON-EPS dataset.

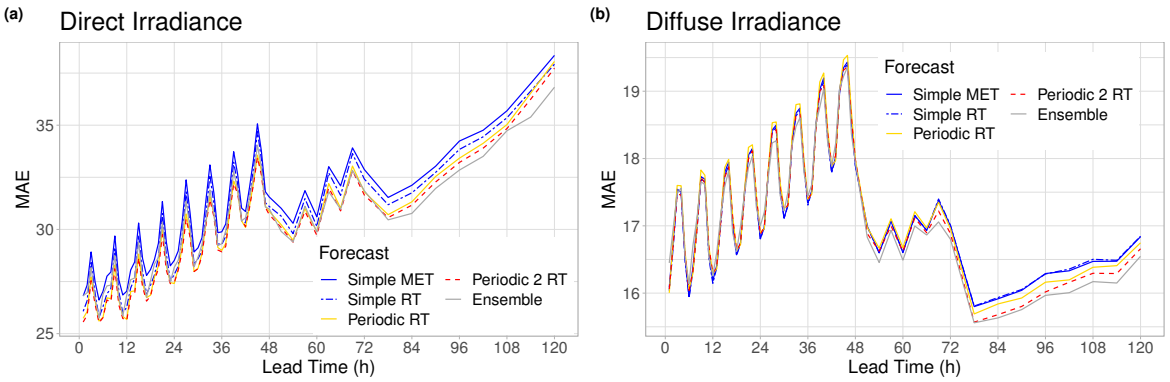

Figure 5.23: MAE of the median forecasts of direct (a) and diffuse (b) irradiance.

evaluated for two datasets. Even though the datasets cover distinct geographical regions, NWP systems, types of solar irradiance and temporal resolutions, the results presented in this section indicate that the proposed postprocessing models are able to consistently and significantly improve the forecast performance of the raw ensemble predictions up to lead times of at least 48 hours. The improvements from postprocessing are larger for the AROME-EPS dataset, possibly due to a lower skill of the raw ensemble predictions resulting from a bias in addition to the observed underdispersion. For the ICON-EPS dataset, we observed that more complex postprocessing models tend to show better predictive performances, but the differences between model variations rarely show a high level of statistical significance. For the GHI predictions of the AROME-EPS dataset, we only found negligible differences when comparing postprocessing models for GHI and CSI. This is in line with the results reported in Yang (2020b) and suggests that the standard practice of normalizing the irradiance forecasts by clear-sky irradiance does not lead to improvements in forecast performance here.

The overall level of improvements achieved via statistical postprocessing of the solar irradiance forecasts of the raw ensemble are comparable to meteorological variables such as precipitation accumulation (Scheuerer, 2014; Baran and Nemoda, 2016) or total cloud cover (Baran et al., 2021a) in case of the ICON-EPS dataset, and slightly larger for the AROME-EPS data. Postprocessing ensemble predictions of those variables is often seen as a more difficult task compared to variables such as temperature (Gneiting et al., 2005) or wind speed (Thorarinsdottir and Gneiting, 2010) for which substantially larger improvements can be achieved, such as in case of the postprocessing on the KIT-Weather portal (e.g., Figure 5.6). Nonetheless, the observed improvements are statistically significant for lead times of up to two days, and will likely be of relevance for solar energy forecasting in terms of potential economic benefits and improved balancing of demand and supply for integrating volatile PV power systems into the electrical grid.

5.3 Wind gust forecasting over Germany

Following the application of the basic EMOS approach in the two preceding case studies, we will apply the full range of statistical postprocessing methods presented in Chapter 4 to a comprehensive dataset for wind gust prediction. The third and last case study provides a systematic comparison of the three groups of postprocessing methods with a focus on the most sophisticated NN-based approaches and builds the foundation for the investigation of feature-dependent postprocessing in Chapter 6. Again, we start with a short introduction to probabilistic wind gust forecasting before going through the three parts of the case study.

Wind gusts are among the most significant natural hazards in central Europe. Accurate and reliable forecasts are therefore critically important to issue effective warnings and protect human life and property. However, as indicated in Section 4.1, wind gusts are a challenging meteorological target variable as they are driven by small-scale processes and local occurrence, so that their predictability is limited even for NWP models run at convection-permitting resolutions. Despite their importance for severe weather warnings, much recent work on ensemble postprocessing has instead focused on temperature and precipitation. Therefore, our overarching aim is to provide a comprehensive review and systematic comparison of the methods for ensemble postprocessing specifically tailored to wind gusts.

Many of the postprocessing methods described in Chapter 4 have been applied for wind speed prediction, but previous work on wind gusts is scarce. Our case study is based on the work of Pantillon et al. (2018), one of the few exceptions, who use a simple EMOS model for postprocessing to investigate the predictability of wind gusts with a focus on European winter storms. They find that although postprocessing improves the overall predictive performance, it fails in cases that can be attributed to specific mesoscale structures and corresponding wind gust generation mechanisms. As a first step towards the development of more sophisticated methods, we adapt existing as well as novel techniques for statistical postprocessing of wind gusts and conduct a systematic comparison of their predictive performance.

5.3.1 Data

Our case study is based on the same dataset as Pantillon et al. (2018) and we refer to their Section 2.1 for a detailed description. The forecasts are generated by the COSMO-DE-EPS mentioned in Section 4.1. The 20-member ensemble is based on initial and boundary conditions from four different global models paired with five sets of physical perturbations (Pantillon et al., 2018). In the following, we will refer to the four groups corresponding to the global models, which consist of five members each, as subensembles. We consider forecasts that are initialized at 00 UTC with a range from 0 to 21 hours. The data ranges from 9 December 2010 when the EPS started in preoperational mode, to the end of 2016, that is, a period of around six years.

Table 5.6: Overview of the available predictors for the COSMO-DE-EPS data. For the meteorological variables, ensemble forecasts are available, with the term '500–1,000 hPa' referring to the specific model levels at 500, 700, 850, 950 and 1,000 hPa. Spatial predictors contain station-specific information.

Feature	Unit or type	Description
		Meteorological variables
VMAX	m/s	Maximum wind, i.e., wind gusts (10 m)
U	m/s	U component of wind (10 m, 500–1,000 hPa)
V	m/s	V component of wind (10 m, 500–1,000 hPa)
WIND	m/s	Wind speed, derived from U and V via $(U^2 + V^2)^{1/2}$ (10 m, 500–1,000 hPa)
OMEGA	Pa/s	Vertical velocity (pressure) (500–1,000 hPa)
T	K	Temperature (ground level, 2 m, and 500–1,000 hPa)
TD	K	Dew point temperature (2 m)
RELHUM	%	Relative humidity (500–1,000 hPa)
TOT_PREC	kg/m^2	Total precipitation (accumulation)
RAIN_GSP	kg/m^2	Large scale rain (accumulation)
SNOW_GSP	kg/m^2	Large scale snowfall – water equivalent (accumulation)
W_SNOW	kg/m^2	Snow depth water equivalent
W_SO	kg/m^2	Column integrated soil moisture (multilayers; 1, 2, 6, 18, 54)
CLCT	%	Total cloud cover
CLCL	%	Cloud cover (800 hPa – soil)
CLCM	%	Cloud cover (400–800 hPa)
CLCH	%	Cloud cover (000–400 hPa)
HBAS_SC	m	Cloud base above mean sea level, shallow connection
HTOP_SC	m	Cloud top above mean sea level, shallow connection
ASOB_S	W/m^2	Net short wave radiation flux (at the surface)
ATHB_S	W/m^2	Net long wave radiation flux (m) (at the surface)
ALB_RAD	%	Albedo (in shortwave)
PMSL	Pa	Pressure reduced to mean sea level
FI	m^2/s^2	Geopotential (500–1,000 hPa)
		Other predictors
yday	Temporal	Cosine transformed day of the year: $\cos[2\pi(t-1)/365]$, where t is the day of the year.
lat	Spatial	Latitude of the station.
lon	Spatial	Longitude of the station.
alt	Spatial	Altitude of the station.
orog	Spatial	Difference of station altitude and model surface height of nearest grid point.
loc_bias	Spatial	Mean bias of wind gust ensemble forecasts from 2010–2015 at the station.
loc_cover	Spatial	Mean coverage of the wind gust ensemble forecasts from 2010–2015 at the station.

In addition to wind gusts, ensemble forecasts of several other meteorological variables generated by the COSMO-DE-EPS are available. Table 5.6 gives an overview of the 61 meteorological variables as well as additional temporal and spatial predictors derived from station information. The forecasts are evaluated at 175 SYNOP stations in Germany operated by DWD, for which hourly observations are available. For the comparison with station data, forecasts from the nearest grid point are taken.

The postprocessing methods presented in Chapter 4 are trained on a set of past forecast-observation pairs in order to correct the systematic errors of the ensemble predictions. Even though many studies are based on rolling training windows consisting of the most recent days only, we will use a static training period. This is common practice in the operational use of

Table 5.7: Overview of the main characteristics of the different postprocessing methods. The number of parameters refers to one trained model instance. The number of models refers to the number of trained model instances per lead time. In case of the NN-based methods, the 10 trained model instances are aggregated to a final forecast.

Method	Estimation	# Parameters	# Models	Local	Seasonal	Forecast distribution
EMOS	CRPS	4	2,100	✓	✓	Truncated logistic
MBM	CRPS	10	2,100	✓	✓	Ensemble
IDR	CRPS	-	175	✓	-	Empirical
EMOS-GB	MLE	248	175	✓	-	Truncated logistic
QRF	Custom	-	175	✓	-	Set of quantiles
DRN	CRPS	9,090	10	-	-	Truncated logistic
BQN	QS	8,013	10	-	-	Bernstein quantile fct.
HEN	MLE	9,684	10	-	-	Piecewise uniform

postprocessing models (Hess, 2020) and can be motivated by studies suggesting that using long archives of training data often lead to superior performance, irrespective of potential changes in the underlying NWP model or the meteorological conditions (Lang et al., 2020). Therefore, we will use the period of 2010–2015 as training set and 2016 as independent test set. The implementation of most of the methods requires the choice of a model architecture and the tuning of specific hyperparameters. To avoid overfitting in the model selection process, we further split the training set into the period of 2010–2014 which is used for training, and use the year 2015 for validation purposes. After finalizing the choice of the most suitable model variant based on the validation period, the entire training period from 2010–2015 is used to fit that model for the final evaluation on the test set.

5.3.2 Model configurations

This section introduces the configuration of the postprocessing methods that are systematically compared for ensemble forecasts of wind gusts. First, note that for each of the postprocessing methods, we fit a separate model for each lead time based on training data consisting only of cases corresponding to that lead time. Table 5.7 gives an overview of the main characteristics of the different postprocessing methods, Table 5.8 of the predictors used and Table 5.9 lists the chosen hyperparameters.

If the evaluation or parts of it are based on a set of quantiles, we generate 125 equidistant quantiles for each test sample. This number is chosen such that the median as well as the quantiles at the levels of a PI with a nominal coverage corresponding to a 20-member ensemble (90.48%) are included and such that the forecast distribution is given by a sufficiently large number of quantiles. The quantiles are then evaluated analogously to an ensemble forecast.

Table 5.8: Overview of the predictors used in the different postprocessing methods. The column "Statistics" comprises the use of any summary statistic derived from the wind gust ensemble (mean, standard deviation, or mean difference).

	Wind gust ensemble		Ensembles of other meteorological variables		Other predictors	
Method	Statistics	Members	Mean	Std. dev.	Temporal	Spatial
EMOS	✓	-	-	-	-	-
MBM	✓	✓	-	-	-	-
IDR	-	✓	-	-	-	-
EMOS-GB	✓	-	✓	✓	✓	-
QRF	✓	-	✓	-	✓	-
DRN	✓	-	✓	-	✓	✓
BQN	-	✓	✓	-	✓	✓
HEN	✓	-	✓	-	✓	✓

EMOS

As described in Section 4.2.1, our EMOS model for wind gusts uses a truncated logistic distribution based on the parameterization described in equations (4.2) and (4.3). Note that we do not specifically account for the existence of subensembles in the parameterization of the distribution parameters, since initial experiments suggested a degradation of predictive performance. We here estimate the parameters by minimizing the CRPS, for which we observed similar results to MLE. In the gradient-based optimization, we proceed analogous to the EMOS models for the KIT-Weather portal and use the L-BFGS-B method to determine the EMOS parameter vector (a, b, c, d).

The EMOS coefficients are estimated locally with a seasonal training scheme where a training set consists of all forecast cases of the previous, current and next month with respect to the date of interest. This results in 12 different training sets for each station, one for each month, that enable an adaption to seasonal changes. In accordance with the results in Lang et al. (2020), this seasonal approach outperforms both a rolling training window as well as training on the entire set.

EMOS-GB

To ensure comparability with the basic EMOS approach, we employ a truncated logistic distribution for the probabilistic forecasts. The parameters are determined using MLE, which resulted in superior predictive performance based on initial tests on the validation data relative to minimum CRPS estimation. We use the ensemble mean and standard deviation of all meteorological variables in Table 5.6 as inputs to the EMOS-GB model. Note that in contrast to the other advanced postprocessing methods introduced below, we here include the

Table 5.9: Final hyperparameter configurations of the models described in Section 5.3.2. The NN-specific hyperparameters are displayed separately in Table 5.10.

Method	Hyperparameter	Value
IDR	Number of subsamples	100
	Subsample ratio	0.5
	Maximum number of iterations (max_iter)	1,000
	Absolute threshold (eps_abs)	0.001
	Relative threshold (eps_rel)	0.001
EMOS-GB	Number of maximum iterations	1,000
	Step size	0.05
	Stopping criterion	AIC
QRF	Number of trees	1,000
	Ratio of predictors considered at each split	0.5
	Minimal node size	5
	Maximal tree depth	20
BQN	Degree of Bernstein polynomials (d in equation (2.33))	12
	Number of equidistant quantiles in QS (n_q in equation (4.26))	99
HEN	Number of (nonequidistant) bins (N in equation (2.40))	20

standard deviation of all variables as potential predictors since we found this to improve the predictive performance. Further, we include the cosine-transformed day of the year in order to adapt to seasonal changes, since a seasonal training approach as applied for EMOS leads to numerically unstable estimation procedures and degraded forecast performance. Although spatial predictors can in principle be included in a similar fashion, we estimate EMOS-GB models locally since we found this approach to outperform a joint model for all stations by a large margin. Our implementation of EMOS-GB is based on the `crch` package (Messner et al., 2016).

MBM

The MBM approach assumes the exchangeability of the ensemble members, but this is in practice not always the case. As described in Section 5.3.1, the COSMO-DE-EPS is based on different NWP submodels resulting in four subensembles. When we apply the general MBM approach based on equations (4.12) and (4.14) to the wind gust data, the existence of the subensembles results in systematic deviations from calibration, especially for a lead time of 0 hours, as the verification rank histograms in Figure 5.24 show.

Due to these systematic deviations from calibration, we extend the MBM approach towards ensembles that exhibit a subensemble structure. In general, let K be the number of the subensembles. Then, we can identify each ensemble member with its associated subensemble

via the function $k : \{1, \ldots, m\} \to \{1, \ldots, K\}$, count the size of each subensemble via

$$m_k := \sum_{i=1}^{m} \mathbb{1}\{k(i) = k\}, \quad k = 1, \ldots, K, \tag{5.9}$$

and define the subensemble mean $\overline{x}_k$ and subensemble mean difference $\delta_k(x)$ via

$$\overline{x}_k := \frac{1}{m_k} \sum_{i=1}^{m} x_i \cdot \mathbb{1}\{k(i) = k\}, \quad k = 1, \ldots, K, \tag{5.10}$$

$$\delta_k(x) := \frac{1}{m_k^2} \sum_{i,l=1}^{m} |x_i - x_l| \cdot \mathbb{1}\{(k(i) = k) \wedge (k(l) = k)\}, \quad k = 1, \ldots, K. \tag{5.11}$$

Using this, we can incorporate the submodel structure by modifying the MBM equations (4.12) and (4.14) to

$$\tilde{x}_i := \left(a + b_{k(i)}\overline{x}_{k(i)}\right) + \left(c + \frac{d_{k(i)}}{\delta_{k(i)}(x)}\right)\left(x_i - \overline{x}_{k(i)}\right), \quad i = 1, \ldots, m, \tag{5.12}$$

where $a, b_1, \ldots, b_K, c, d_1, \ldots, d_K \in \mathbb{R}$ are the MBM parameters. While the subensembles share the intercept parameters a and c, the members of each subensemble are adjusted separately based on the corresponding subensemble mean and subensemble mean difference using the parameters $b_1, \ldots, b_K, d_1, \ldots, d_K$. Parameter estimation can be carried out analogously to the general MBM approach via MLE or CRPS estimation.

For the COSMO-DE-EPS, we have $K = 4$ with the identification function $k(i) = \lceil i/5 \rceil$ for $i = 1, \ldots, 20$ and the MBM parameter vector $(a, b_1, \ldots, b_4, c, d_1, \ldots, d_4)$. This increases the number of parameters by six, but substantially improves performance and mostly eliminates the relicts of the subensemble structure, as the PIT histograms in Figure 5.25 exemplify. One downside of this modification is that it increases the computational time drastically, here by a factor of around 17. As alternative modifications to equation (5.12), we also applied MBM separately to each subensemble with a separate parameter c for each submodel, which performs equally well but is more complex. Further, we tested leaving out the parameter d, both with a single c and $c_1, \ldots, c_4$ for the subensembles, but this resulted in a worse predictive performance still exhibiting systematic deviations in the histograms.

The scoring rule of choice is the CRPS, as MLE resulted in less well-calibrated forecasts with slightly worse overall performance. The training is performed analogously to EMOS, utilizing a local and seasonal training scheme. In particular, accounting for potential seasonal changes via seasonal training substantially improved performance compared to using the entire available training set.

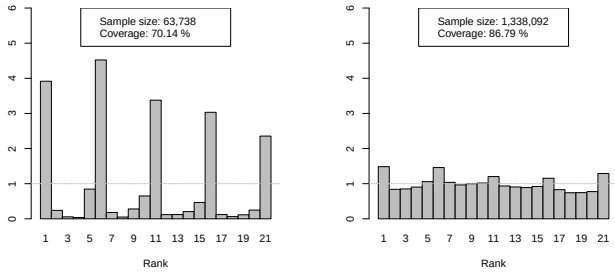

Figure 5.24: Verification rank histograms of MBM forecasts based on equation (4.12) with CRPS estimation on the entire training set for a lead time of 0 hours (left) and greater than 0 hours (right). Coverage refers to the empirical coverage of a PI with a nominal coverage corresponding to a 20-member ensemble (90.48%).

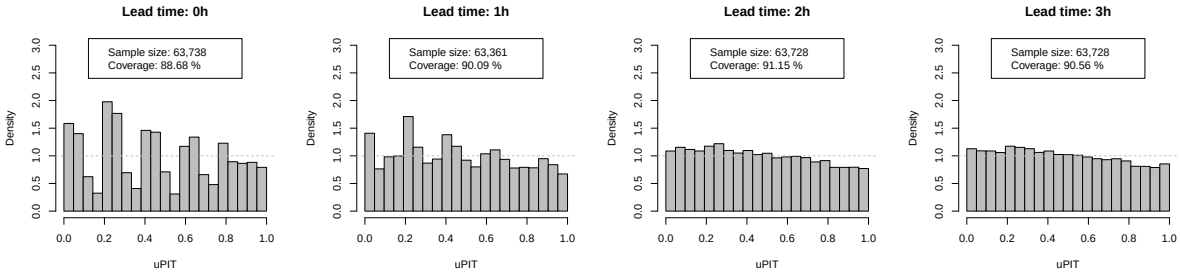

Figure 5.25: PIT histograms of the MBM forecasts for lead times from 0 to 3 hours (left to right) over all stations. Coverage refers to the empirical coverage of a PI with a nominal coverage corresponding to a 20-member ensemble (90.48%).

IDR

The only implementation choice required for IDR is the selection of a partial order on the covariate space. Among the choices presented in Section 4.2.4, the empirical stochastic order and the empirical increasing convex order are the two appropriate choices for the situation at hand, as we use all ensemble members as predictors. We selected the empirical stochastic order which resulted in slightly better results on the validation data. We further considered an alternative model formulation where only the ensemble mean was used as predictor, which reduces to the special case of a less complex distributional (single) index model (Henzi et al., 2021a), but did not improve predictive performance.

We implement IDR as a local model, treating each station separately since it is not obvious how to incorporate station-specific information into the model formulation. Given the limited amount of training data available, we further consider only the wind gust ensemble as predictor variable. Following suggestions of Henzi et al. (2021b), we use subsample aggregation (subbagging) and apply IDR on 100 random subsamples half the size of the available training set. IDR is implemented using the `isodistrreg` package (Henzi et al., 2019).

QRF

In contrast to EMOS-GB, only the ensemble mean values of the additional meteorological variables are integrated as predictor variables, since we found that including the standard deviations as well led to more overdispersed forecasts and degraded forecast performance. A potential reason is given by the random selection of predictors at each node, which limits the automated selection of relevant predictors in that a decision based on a subset of irrelevant predictors only may lead to overfitting (Hastie et al., 2009, Section 15.3.4).

Although spatial predictors can be incorporated into a global, joint QRF model for all stations generating calibrated forecasts, we found that the extant practice of implementing local QRF models separately at each station (Taillardat et al., 2016; Rasp and Lerch, 2018) results in superior predictive performance and avoids the increased computational demand both in terms of required calculations and memory of a global QRF variant (Taillardat and Mestre, 2020). Our implementation is based on the `ranger` package (Wright and Ziegler, 2017).

Neural networks

To determine hyperparameters of the NN models such as the learning rate, the embedding dimension or the number of nodes in a hidden layer, we perform a two-step, semiautomated hyperparameter tuning based on the validation set. In an automated procedure, we first find a small number of hyperparameter sets that perform best for an individual network, then we manually select that set that yields the best aggregated forecasts. Overall, the results are relatively robust to a wide range of tuning parameter choices, and we found that increasing the number of layers or the number of nodes in a layer did not improve predictive performance. Relative to the models used in Rasp and Lerch (2018), we increased the embedding dimension and used a softplus activation function. The exact configuration slightly varies across the three model variants introduced in the following. In addition to the station embedding, the spatial features in Table 5.6, and the temporal predictor, we found that including only the mean values of the meteorological predictors, but not the corresponding standard deviations, improved the predictive performance. These results are in line with those of QRF and those of Rasp and Lerch (2018) who find that the standard deviations are only of minor importance for explaining and improving the NNs predictions. Before fitting the NN models, each predictor variable was normalized by subtracting the mean value and dividing by the standard deviation based on the training set excluding the validation period. All NN models were implemented via the `R` (R Core Team, 2021) interface to `keras` (2.4.3; Allaire and Chollet, 2020) built on `tensorflow` (2.3.0; Allaire and Tang, 2020). The network architectures and the neural network-specific hyperparameter are displayed in Table 5.10.

Table 5.10: Overview of the configuration of the individual networks in the NN-based methods.

Hyperparameter	DRN	BQN	HEN
Learning rate	$5 \cdot 10^{-4}$	$5 \cdot 10^{-4}$	$5 \cdot 10^{-4}$
Epochs	150	150	150
Patience	10	10	10
Batch size	64	64	64
Embedding dimension	10	10	10
Hidden layers	2	2	2
Nodes per layer	(64, 32)	(48, 24)	(64, 32)
Activation	Softplus	Softplus	Softplus
Output nodes	2	13	20
Output activation	Softplus	Softplus	Softmax
Size of network ensemble	10	10	10

DRN

We adapt DRN to wind gust forecasting by using a truncated logistic distribution. In contrast to the Gaussian predictive distribution used by Rasp and Lerch (2018), this leads to additional technical challenges due to the truncation, i.e., the division by $1 - F(0; \mu, \sigma)$, which induces numerical instabilities. To stabilize training, we enforce $\mu \geq 0$ by applying a softplus activation function in the output nodes, resulting in $1 - F(0; \mu, \sigma) \geq 0.5$. As noted for the EMOS model of wind speed for the KIT-Weather portal in Section 5.1.2, the restrictions of the parameter space can be considered to be negligible. The optimum score estimation based on the LogS and CRPS yields similar results with the CRPS performing slight better, hence we use CRPS estimation.

BQN

The BQN loss function in equation (4.26) is based on $n_q = 99$ equidistant quantiles corresponding to steps of 1%. Regarding the degree of the Bernstein polynomials, Bremnes (2020) considers a degree of $d = 8$. We found that increasing the degree to 12 resulted in better calibrated forecasts and improved predictive performance on the validation data. Again following Bremnes (2020) and in contrast to our implementation of DRN and HEN, we use all 20 ensemble member forecasts of wind gust sorted with respect to the predicted speed as input instead of the ensemble mean and standard deviation.

HEN

For the application to wind gusts, we found that the binning scheme is an essential factor, e.g., a fine equidistant binning of length 0.5 m/s leads to physically inconsistent forecasts. Based

on initial experiments on the validation data, we devise a data-driven binning scheme starting from one bin per unique observed value and merging bins to end up with a total number of $N = 20$. We start with one bin for each observation (which only take a certain amount of values for reporting reasons) and merge the bin that contains the least amount of observations with the smaller one of the neighbouring bins. We additionally put constraints on the bins. The first bin should have a length of at most 2 m/s, the last at most 7 m/s and the others at most 5 m/s. In the aggregation procedure, the binning in terms of the probabilities is reduced to a minimal bin width of 0.01% for numerical reasons. In the output nodes, we apply a softmax function to ensure that the obtained probabilities sum to 1. Following the suggestion in Section 4.3.4, we estimate the network parameters using the categorical cross-entropy, which corresponds to MLE.

5.3.3 Results

In this section, we evaluate the predictive performance of the postprocessing methods based on the test period that consists of all data from 2016. Since we considered forecasts from one initialization time only, systematic changes over the lead time are closely related to the diurnal cycle.

Predictive performance of the COSMO-DE-EPS and a climatological baseline

The predictive performance of the EPS coincides with findings in Sections 5.1 and 5.2 in that the ensemble predictions are biased and strongly underdispersed, see Figure 5.26 for the corresponding verification rank histograms. We here highlight two peculiarities of the EPS. The first is the so-called *spin-up effect* (e.g., Kleczek et al., 2014), which refers to the time the numerical model requires to adapt to the initial and boundary conditions and to produce structures consistent with the model physics. This effect can not only be seen in the verification rank histograms in Figure 5.26, where we observe a clear lack of ensemble spread in the 0 hour forecasts within each of the four subensembles and only a small spread between them, but also for the ensemble range and the bias of the ensemble median prediction displayed in Figure 5.27, where a sudden jump at the 1 hour forecasts occurs.

The temporal development of the bias and ensemble range shown in Figure 5.27 indicates another meteorological effect, the *evening transition* of the planetary boundary layer. When the sun sets, the surface and low-level air that have been heated over the course of the day cool down and thermally driven turbulence ceases. This sometimes quite abrupt transition to calmer, more stable conditions strongly affects the near-surface wind fields subject to the local conditions (e.g., Mahrt, 2017). For lead times up to 18 hours (corresponding to 19 local time in winter and 20 in summer), the ensemble range increases together with an improvement in calibration. However, at the transition in the evening, the overall bias increases and the

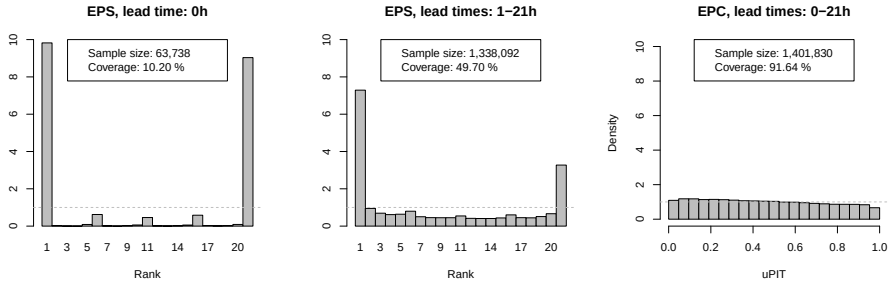

Figure 5.26: Verification rank histograms of 0 and 1–21 hour forecasts of the COSMO-DE-EPS and PIT histogram of the EPC forecasts over all lead times for all stations (left to right). Coverage refers to a PI with a nominal coverage corresponding to a 20-member ensemble (90.48%).

calibration becomes worse for most stations indicating increasing systematic errors. This could be related, for example, to the misrepresentation of the inertia of large eddies in the model or errors in radiative transfer at low sun angles.

In addition to the raw ensemble predictions, we further consider a climatological reference forecast as a benchmark method. The *extended probabilistic climatology* (EPC; Vogel et al., 2018; Walz et al., 2021) is an ensemble based on past observations considering only forecasts at the same time of the year. We create a separate climatology for each station and hour of the day that consists of past observations from the previous, current and following month around the date of interest. The observational dataset ranges back to 2001, thus EPC is built on a dataset of 15 years. Not surprisingly, EPC is well-calibrated (Figure 5.26). However, it shows a minor positive bias, which is likely due to the generally lower level of wind gusts observed in 2016 compared to the years on which EPC is based, which is illustrated in Figure 5.28.[16]

Comparison of the postprocessing methods

Figure 5.29 shows PIT histograms for all postprocessing methods. All approaches substantially improve the calibration compared to the raw ensemble predictions and yield well-calibrated forecasts, except for IDR which results in underdispersed predictions. The PIT histograms of the parametric methods based on a truncated logistic distribution (EMOS, EMOS-GB and DRN) all exhibit similar minor deviations from uniformity caused by a lower tail that is too heavy.[17] The semi- and nonparametric methods MBM, QRF, BQN and HEN are all slightly skewed to the left, in line with the histogram of EPC. Further, we observe a minor overdispersion for the QRF forecasts.

Table 5.11 summarizes the values of proper scoring rules and other evaluation metrics to compare the overall predictive performance of all methods. While the ensemble predictions outperform the climatological benchmark method, all postprocessing approaches lead to

[16] See Wohland et al. (2018) for a discussion of the 2016 wind variability in the context of wind power generation.
[17] Analogous to the EMOS forecasts for wind speed in Section 5.1, which are also based on a zero-truncated logistic distribution.

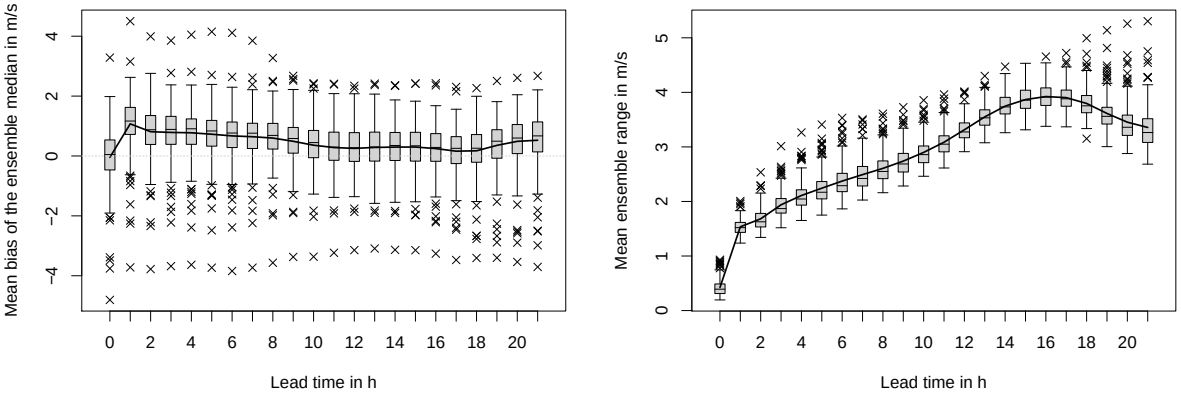

Figure 5.27: Boxplots of the stationwise mean bias of the ensemble median (left) and the mean ensemble range (right) of the COSMO-DE-EPS forecasts as functions of the lead time. The black line indicates the average over all samples.

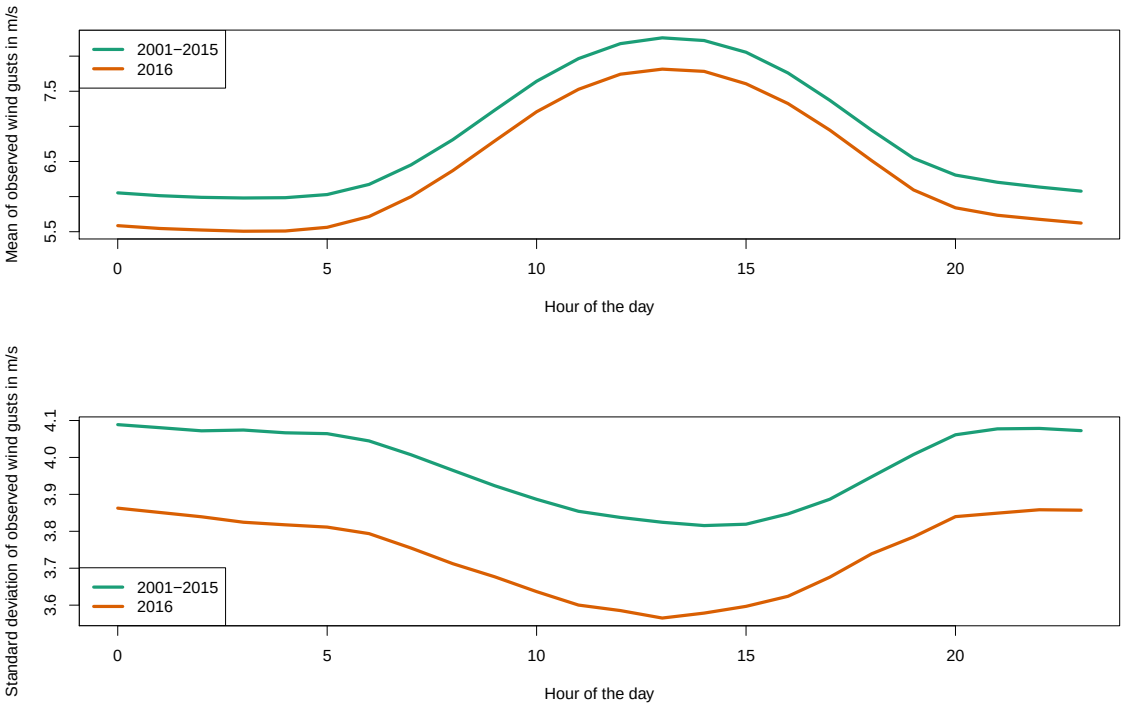

Figure 5.28: Mean (top) and standard deviation (bottom) of the observed wind gusts dependent on the hour of the day in the years 2001–2015 (green) and 2016 (orange)

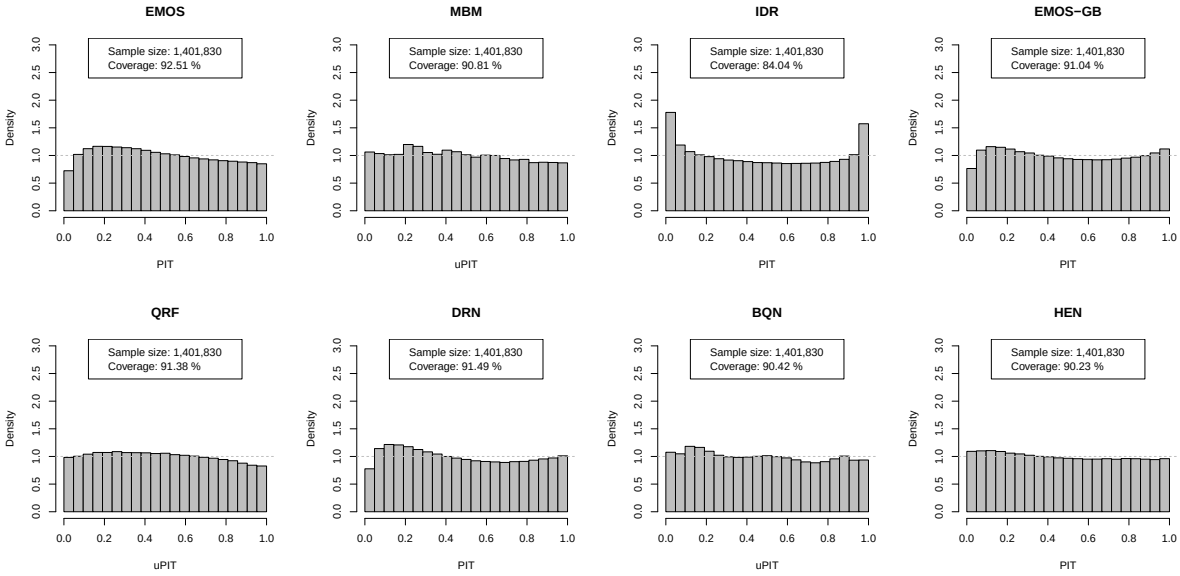

Figure 5.29: PIT histograms of all postprocessing methods, aggregated over all lead times and stations. Coverage refers to the empirical coverage of a PI with a nominal coverage corresponding to a 20-member ensemble (90.48%).

substantial improvements. Among the different postprocessing methods, the three groups of approaches introduced in Chapter 4 show systematic differences in their overall performance. In terms of the CRPS, the basic methods already improve the ensemble by 26–29%. Incorporating additional predictors via the ML methods further increases the skill, where the NN-based approaches, in particular DRN and BQN, perform best. The MAE and RMSE lead to analogous rankings, and all methods clearly reduce the bias of the EPS. Among the well-calibrated postprocessing methods, the NN-based methods yield the sharpest forecast distributions, followed by QRF, EMOS-GB and the basic methods. Thus, we conclude that the gain in predictive performance is mainly based on an increase in sharpness. Overall, BQN results not only in the best coverage, but also in the sharpest PIs.

We further consider the total computation time required for training the postprocessing models. However, note that a direct comparison of computation times is difficult due to the differences in terms of software packages and parallelization capabilities. Not surprisingly, the simple EMOS method was the fastest with only 19 minutes. The network-based methods were not much slower than QRF and faster than EMOS-GB, which is based on almost twice as much predictors as the other advanced methods what approximately doubled the computational costs. MBM here is an extreme outlier and requires a computation time of over 35 days in total, in particular due to our adaptations to the subensemble structure discussed in Section 5.3.2.

Table 5.11: Evaluation metrics for EPC, COSMO-DE-EPS, and all postprocessing methods averaged over all lead times and stations. The PI length and coverage refer to a PI with a nominal coverage corresponding to a 20-member ensemble (90.48%). The best methods are indicated in bold.

Method	CRPS	MAE	RMSE	Bias	PI length	Coverage	Runtime
EPC	1.72	2.44	3.26	-0.13	10.73	91.64%	-
EPS	1.33	1.63	2.16	0.47	2.85	47.91%	-
EMOS	0.95	1.32	1.80	0.05	5.94	92.51%	**19 min**
MBM	0.97	1.34	1.80	0.04	6.10	90.81%	51,242 min
IDR	0.98	1.36	1.84	0.01	**4.72**	84.04%	8,100 min
EMOS-GB	0.88	1.23	1.69	-0.06	5.24	91.04%	510 min
QRF	0.87	1.22	1.66	-0.03	5.41	91.38%	282 min
DRN	**0.84**	1.18	1.61	0.03	5.05	91.49%	399 min
BQN	0.84	**1.18**	**1.61**	**0.00**	4.94	**90.42%**	387 min
HEN	0.86	1.21	1.64	-0.04	5.07	90.23%	321 min

LEAD TIME-SPECIFIC RESULTS

To investigate the effects of the different lead times and hours of the day on the predictive performance, Figure 5.30 shows various evaluation metrics as function of the lead time. While the CRPS values and the improvements over the raw ensemble predictions (Figures 5.30a,b) show some variations over the lead times, the overall rankings among the different methods and groups of approaches are consistent. In particular, the rankings of the individual postprocessing models remain relatively stable over the day.

The spin-up effect is clearly visible in that the mean bias drastically increases from the 0 to the 1 hour forecasts of the EPS and leads to a worse CRPS despite the increase of the ensemble range (Figure 5.27). The postprocessed forecasts, however, are able to successfully correct the biases induced in the spin-up period while benefiting from the increase in ensemble range. Hence, the CRPS becomes smaller and the skill is the largest through all lead times. Although we improve the MBM forecast by incorporating the submodel structure, the adjusted ensemble forecasts are still subject to systematic deviations from calibration for lead times of 0 and 1 hour, as illustrated in Figure 5.25.

Following the first hours, a somewhat counterintuitive trend can be observed in that the predictive performance of the EPS improves up to a lead time of 10 hours. This is in particular due to improvements in terms of the spread of the ensemble over time. By contrast, the predictive performance of the climatological baseline model is affected more by the diurnal cycle since observed wind gusts and their variability tend to be higher during daytime. The performance of the postprocessing models is neither affected by the increased spread of the EPS, nor by larger gust observations, and slightly decreases over time until the evening

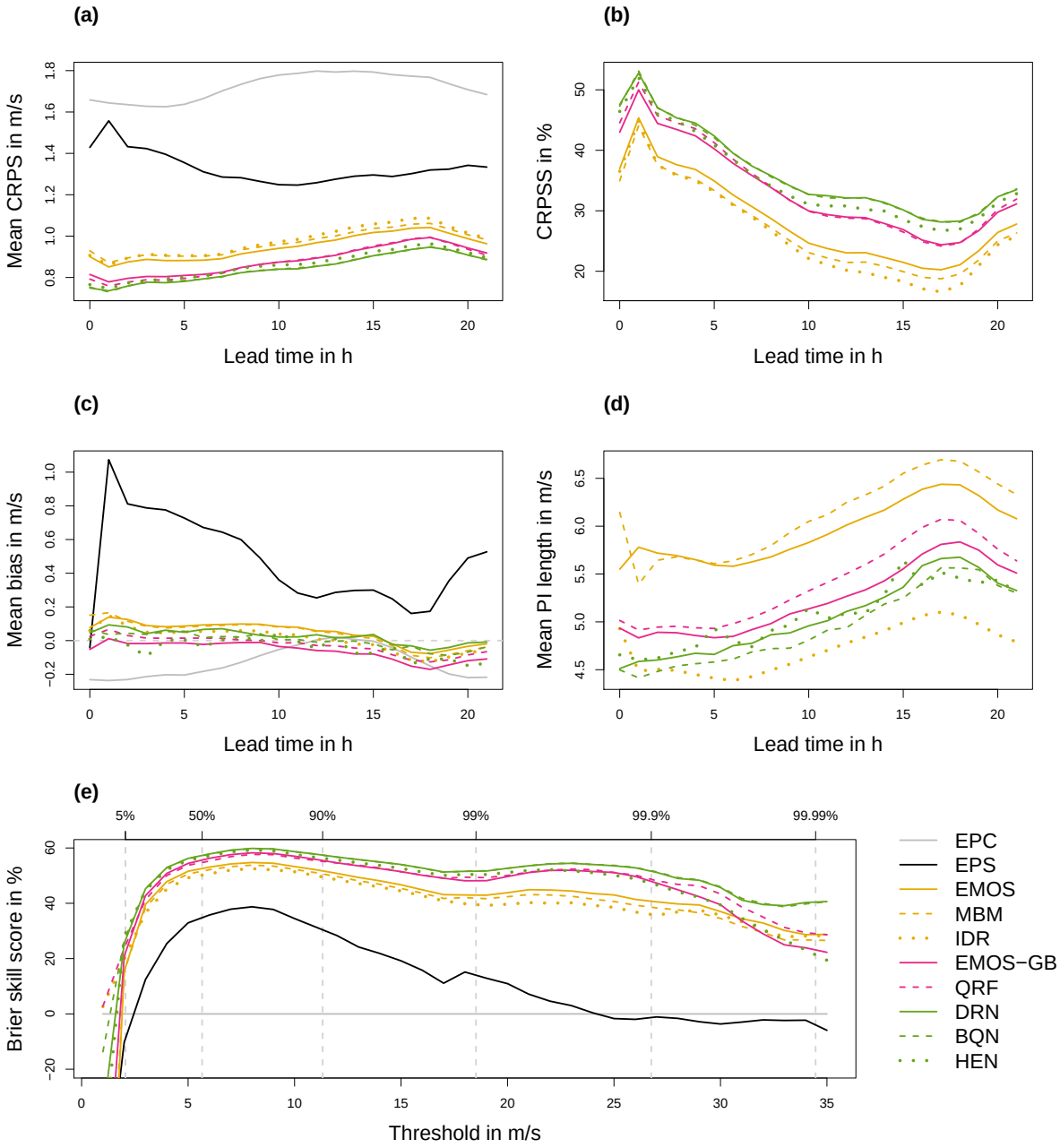

Figure 5.30: Mean CRPS (a), CRPSS with respect to the raw ensemble predictions (b), mean bias (c) and mean PI length (d) of the postprocessing methods as functions of the lead time, averaged over all stations. Panel (e) shows the BSS with respect to the climatological EPC forecast as function of the threshold, averaged over all lead times and stations, where the dashed vertical lines indicate the quantiles of the observed wind gusts at levels given at the top axis.

transition. This is in line with wider PIs that represent increasing uncertainty for longer lead times, while the mean bias and coverage are mostly unaffected.

This general trend changes with the evening transition at a lead time of around 18 hours. The CRPS of the climatological reference model decreases due to a better predictability of the wind gust forecasts. By contrast, the CRPS of the ensemble increases, again driven by an increase in bias and a decrease in spread that comes with a smaller coverage. The numerical model thus appears to not be fully capable of capturing the relevant physical effects, and introduces systematic errors. The bias and coverage of the postprocessing methods do not change drastically, while the PIs of the postprocessing methods become smaller, which is in line with the more stable conditions at nighttime. Therefore, the CRPS of the postprocessing methods becomes better again.

To assess the forecast performance for extreme wind gust events, Figure 5.30e shows the BSS with respect to the climatological EPC forecast as a function of the threshold value, averaged over all stations and lead times. For larger threshold values, the EPS rapidly looses skill and does not provide better predictions than the climatology for thresholds above 25 m/s. By contrast, all postprocessing methods retain positive skill across all considered threshold values. The predictive performance decreases for very high threshold values above 30 m/s, in particular for EMOS-GB and QRF. Note that the EPS and all postprocessing methods besides the analog-based QRF and IDR have negative skill scores for very small thresholds, but this is unlikely to be of relevance for any practical application.

Station-specific results and statistical significance

We further investigate the station-specific performance of the different postprocessing models and in particular investigate whether the locally adaptive networks that are estimated jointly for all stations also outperform the locally estimated methods at the individual stations. Figure 5.31 shows a map of all observation stations indicating the station-specific best model in terms of the CRPS, and demonstrates that at 162 of the 175 stations a network-based method performs best. While none of the basic methods provides the best forecasts at any station, QRF or EMOS-GB perform best at the remaining 13 stations. Most of these station are located in mountainous terrain or coastal regions that are likely subject to specific systematic errors, which might favor a location-specific modeling approach.

Next, we shortly compare the methods in terms of the ES, which is a multivariate evaluation measure and takes therefore spatio-temporal consistency into account. Here, we calculate the ES of the time series forecasts consisting of forecasts for all lead times at a given initialization time, i.e., a set of 22 forecasts corresponding to the lead times 0–21 hours. The computation of the ES is based on the ensemble in case of the EPS and MBM or a sample of 20 randomly drawn realizations from the forecast distribution otherwise. The results are shown in Figure 5.32. We can see that the ranking is closely related to that obtained in Table 5.11, and that the NN methods performed best at 164 of the 175 stations. The main difference is that MBM, which

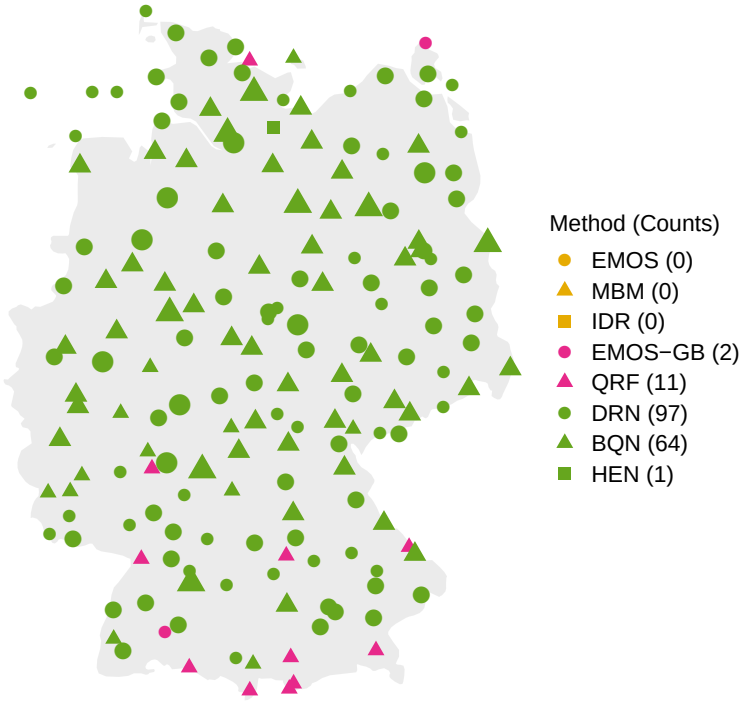

Figure 5.31: Best method at each station in terms of the CRPS, averaged over all lead times. The point sizes indicate the level of statistical significance of the observed CRPS differences compared to the methods only from the other groups of methods for all lead times. Three different point sizes are possible, with the smallest size indicating statistically significant differences for at most 80% of the performed tests, the middle size for up to 95% and the largest 95% or more.

preserves spatio-temporal correlations of the ensemble forecasts, outperforms IDR as well as EMOS, which performs better for univariate measures. Despite that, the ML methods are still superior to MBM and perform best at all but one station, at which MBM performs best.

Last, we evaluate the statistical significance of the differences in the predictive performance in terms of the CRPS between the competing postprocessing methods. To that end, we perform DM tests of equal predictive performance for each combination of station and lead time, and apply a Benjamini-Hochberg procedure to account for potential temporal and spatial dependencies of forecast errors in this multiple testing setting.[18] We find that the observed score differences are statistically significant for a high ratio of stations and lead times (Table 5.12). In particular, DRN and BQN significantly outperform the basic models at

[18] For Table 5.12, we applied the Benjamini-Hochberg correction for each pair of methods separately considering tests for each combination of location and lead time. For Figure 5.31, we applied the correction for each location separately considering the tests comparing the best method with the methods from the other groups for all lead times.

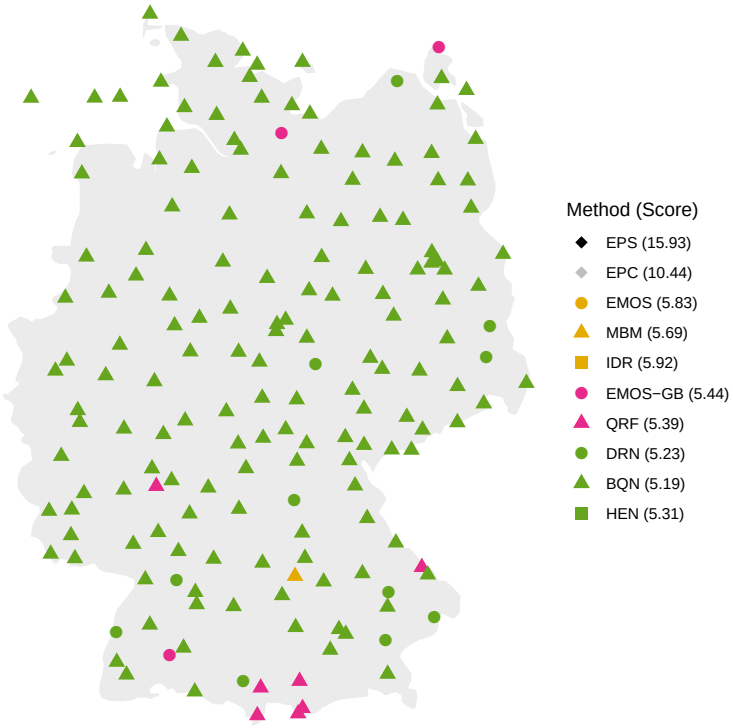

Figure 5.32: Best methods at each station in terms of the ES. The given score denotes the overall ES.

more than 94%, and even significantly outperform QRF and EMOS-GB at more than 50% of all combinations of stations and lead times. Among the locally estimated methods, QRF performs best but only provides significant improvements over the NN-based methods for around 1% of the cases.

To assess station-specific effects of the statistical significance of the score differences, the sizes of the points indicating the best models in Figure 5.31 are scaled by the degree of statistical significance of the results when compared to all models from the two other groups of methods. For example, if DRN performs best at a station, the corresponding point size is determined by the proportion of rejections of the null hypothesis of equal predictive performance at that station when comparing DRN to EMOS, MBM, IDR, EMOS-GB and QRF (but not the other NN-based models) for all lead times in a total of $5 \cdot 22$ DM tests. Generally, if a locally estimated ML approach performed best at one station, the significance tends to be lower than when a network-based method performs best. The most significant differences between the groups of methods can be observed in central Germany, where most stations likely exhibit similar characteristics compared to coastal areas in northern Germany or mountainous regions in southern Germany.

Table 5.12: Ratio of lead time-station combinations (in %) where pairwise DM tests indicate statistically significant CRPS differences after applying a Benjamini-Hochberg procedure to account for multiple testing for a nominal level of $\alpha = 0.05$ of the corresponding one-sided tests. The (i, j)-entry in the ith row and jth column indicates the ratio of cases where the null hypothesis of equal predictive performance of the corresponding one-sided DM test is rejected in favor of the model in the ith row when compared to the model in the jth column. The remainder of the sum of (i, j)- and (j, i)-entry to 100% is the ratio of cases where the score differences are not significant.

	EPC	EPS	EMOS	MBM	IDR	EMOS-GB	QRF	DRN	BQN	HEN
EPC		5.4	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0
EPS	78.9		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
EMOS	99.3	99.9		84.8	51.1	0.0	0.0	0.0	0.0	0.0
MBM	99.3	99.8	0.0		5.7	0.0	0.0	0.0	0.0	0.0
IDR	98.7	99.2	0.0	1.7		0.0	0.0	0.0	0.0	0.0
EMOS-GB	100.0	99.9	69.5	87.5	87.3		0.5	0.2	0.2	1.2
QRF	100.0	99.9	70.3	88.0	91.9	6.1		1.0	1.1	2.7
DRN	99.9	100.0	94.2	97.7	97.3	58.0	52.8		1.8	44.7
BQN	99.9	100.0	94.2	97.3	97.4	56.6	53.1	1.0		43.4
HEN	99.6	99.9	87.0	94.2	93.6	29.6	26.1	0.1	0.0	

Feature importance

The results presented in the previous section demonstrate that the use of additional features improves the predictive performance by a large margin. Here, we assess the effects of the different inputs on the model performance to gain insight into the importance of meteorological variables and better understand what the models have learned. Many techniques have been introduced in order to better interpret ML methods, in particular NNs (McGovern et al., 2019), and we will focus on distinct approaches tailored to the individual ML methods at hand and separately assess the feature importance for the individual methods.

Feature importance for EMOS-GB and QRF

Since the second group of methods relies on locally estimated, separate models for each station, the importance of specific predictors will often vary across different locations and thus make an overall interpretation of the model predictions more involved.

In the case of EMOS-GB, we treat the location and scale parameters separately and consider a feature to be more important the larger the absolute value of the estimated coefficient value is. In the interest of brevity, we here discuss some general properties only. Overall, the interpretation is challenging due to a large variation across stations, in particular, during the spin-up period. In general, the mean value of the wind gust predictions is selected as the most important predictor for the location parameter, followed by other wind-related predictors and

the temporal information about the day of the year (Figure 5.33). For the scale parameter, the standard deviation of the ensemble predictions of wind gust is selected as the most important predictor, followed by the ensemble mean (Figure 5.34). Other meteorological predictors tend to only contribute relevant information for specific combinations of lead times and stations, parts of which might be physically inconsistent coefficient estimates due to random effects in the corresponding datasets. Selected examples are presented in Figure 5.35.

For QRF, we utilize an out-of-bag estimate of the feature importance based on the training set (Breiman, 2001). The procedure is similar to what we apply for the NN-based models below, but uses a different evaluation metric directly related to the algorithm for constructing individual decision trees, see Wright and Ziegler (2017) for details. Figure 5.36 shows the feature importance for some selected predictor variables as a function of the lead time. Interestingly, the 10 most important predictors (two of which are included in Figure 5.36) are variables that directly relate to different characteristics of wind speed predictions from the ensemble. This can be explained by the specific structure of RF models. Since these predictor variables are highly correlated, they are likely to serve as replacements if other ones are not available in the random selection of potential candidate variables for individual splitting decisions. The standard deviation of the wind gust ensemble is only of minor importance during the spin-up period. Besides the wind-related predictors from the EPS, the day of the year, the net shortwave radiation flux prediction as well as the relative humidity prediction at 1,000 hPa are selected as important predictors, particularly for longer lead times corresponding to later times of the day and potentially again indicating an effect of the evening transition. In particular, the shortwave radiation flux indicates the sensitivity of the wind around sunset to the maintenance of turbulence by surface heating, an effect not seen in the morning when the boundary layer grows more gradually.

Feature importance for NN-based methods

To investigate the feature importance for NNs, we follow Rasp and Lerch (2018) and use a permutation-based measure that is given by the decrease in terms of the CRPS in the test set when randomly permuting a single input feature, using the mean CRPS of the model based on unpermuted input features as reference. In order to eliminate the effect of the dependence of the forecast performance on the lead times, we calculate the relative permutation importance.

To introduce the notion of *permutation importance* (Rasp and Lerch, 2018; McGovern et al., 2019), we use ξ to denote the ith predictor ($i = 1, \dots, p$), $F(\boldsymbol{x}_{\cdot j})$ to denote the probabilistic forecast generated by an NN-based postprocessing method based on the jth sample of a test set of size n ($j = 1, \dots, n$), and π to denote a random permutation of the set $\{1, \dots, n\}$.[19] The permutation importance of ξ with respect to the test set $\{(\boldsymbol{x}_{\cdot 1}, y_1), \dots, (\boldsymbol{x}_{\cdot n}, y_n)\}$ and

[19] In general, the permutation importance can be calculated for any postprocessing method and any scoring rule. Here, we focus on the NN-based approaches and the CRPS.

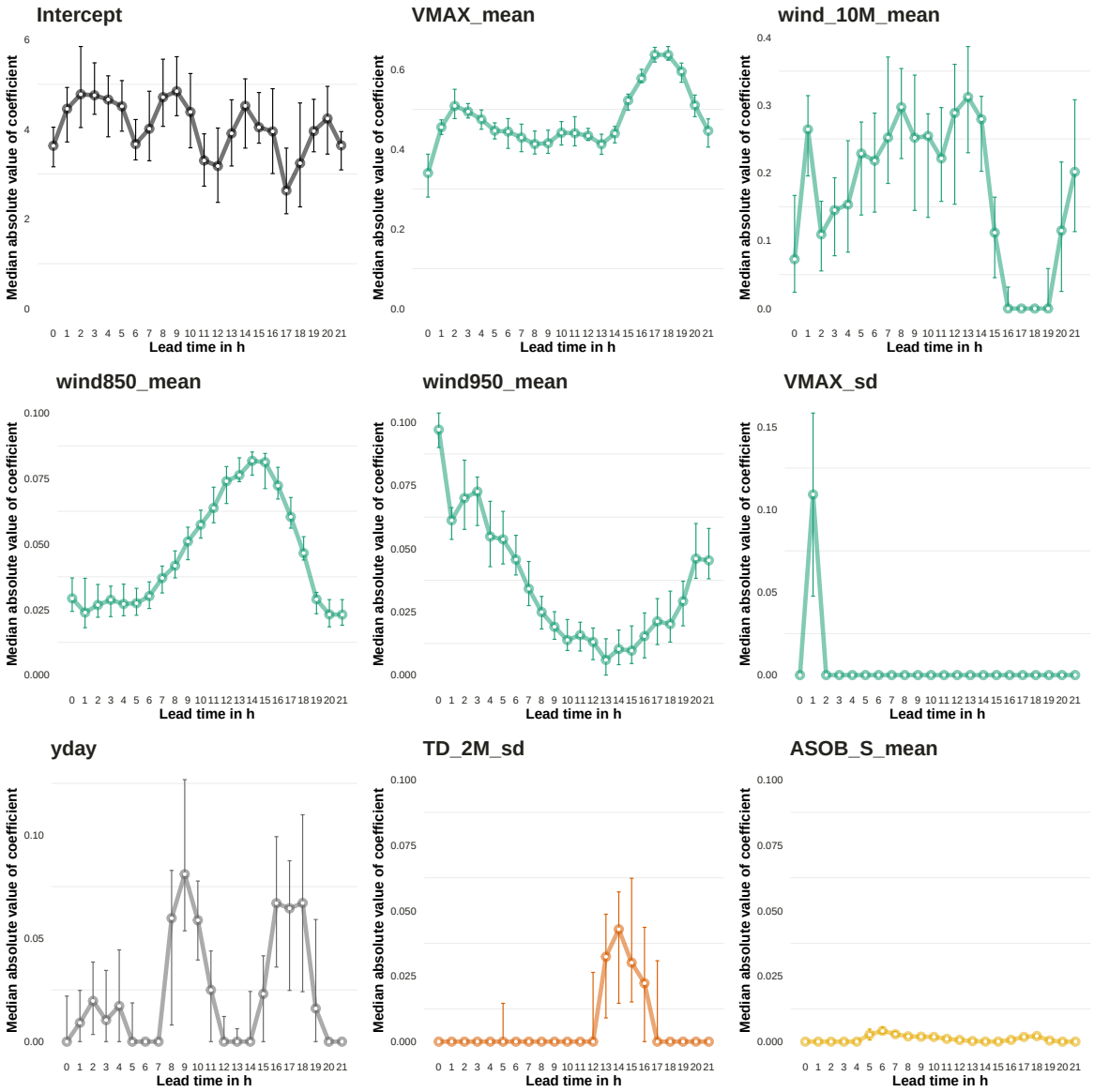

Figure 5.33: Median of stationwise absolute values of the location parameter coefficients for selected predictors (Table 5.6) of the EMOS-GB model as functions of the lead time. The error bars indicate a bootstrapped 95% confidence interval of the median. Note the different scale of the vertical axes.

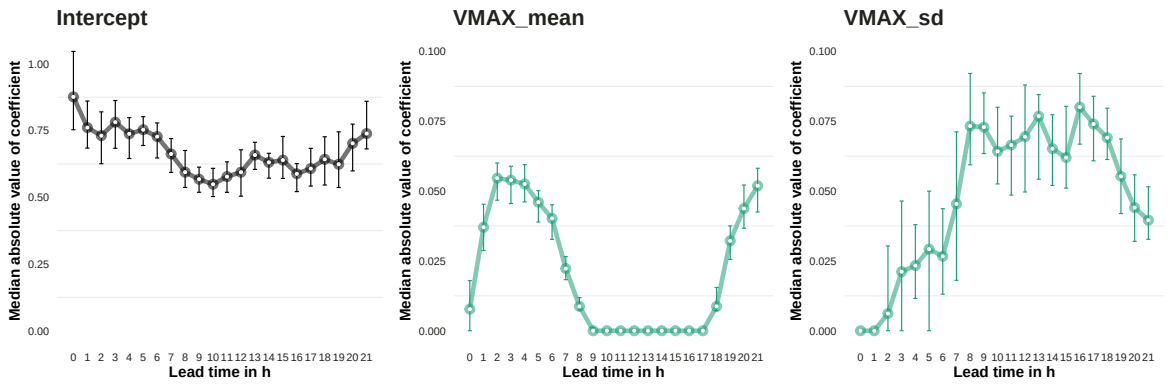

Figure 5.34: Median of stationwise absolute values of the scale parameter coefficients for selected predictors (Table 5.6) of the EMOS-GB model as functions of the lead time. The error bars indicate a bootstrapped 95% confidence interval of the median. Note the different scale of the vertical axes.

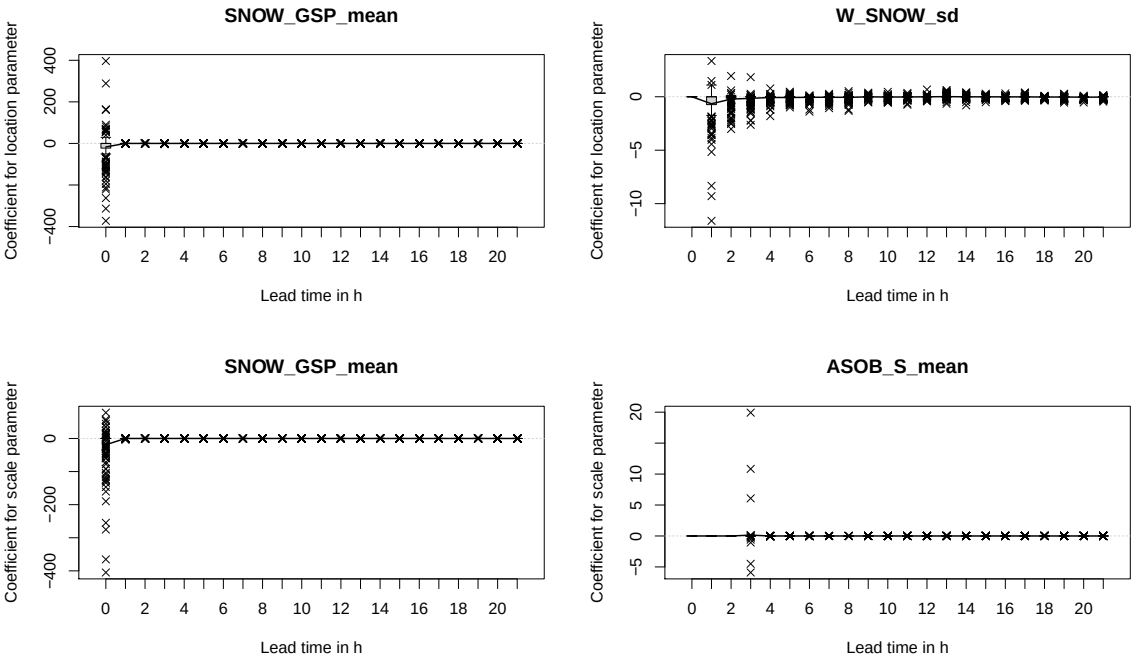

Figure 5.35: Boxplots of stationwise parameter coefficients for selected predictors (Table 5.6) of the EMOS-GB model as functions of the lead time, illustrating severe outliers for some combinations of stations and lead times. The solid line indicates the mean coefficient values averaged over all stations. Note the different scale of the vertical axes.

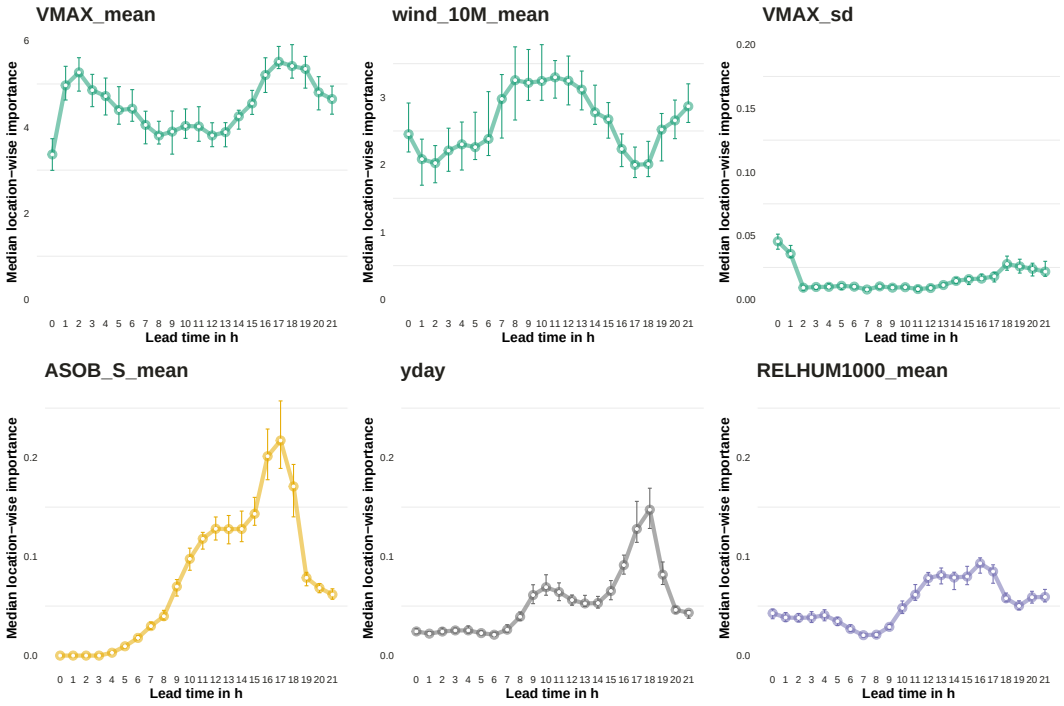

Figure 5.36: Median of stationwise feature importance for selected predictors (Table 5.6) of the QRF model as functions of the lead time. The error bars indicate a bootstrapped 95% confidence interval of the median. Note the different scale of the vertical axes.

permutation π is defined by

$$\Delta\left(\xi;\pi\right) := \frac{1}{n}\sum_{j=1}^{n}\Big(\mathrm{CRPS}(F(\tilde{\boldsymbol{x}}_{\cdot j}^{\pi}), y_j) - \mathrm{CRPS}(F(\boldsymbol{x}_{\cdot j}), y_j)\Big)\,, \tag{5.13}$$

where $\tilde{\boldsymbol{x}}_{\cdot j}^{\pi}$ is the jth sample permuted in ξ with respect to π, which is given by

$$\tilde{\boldsymbol{x}}_{l,j}^{\pi} := \tilde{\boldsymbol{x}}_{l,j}^{\pi}(\xi) := \begin{cases} \boldsymbol{x}_{l,j} & l \neq i, \\ \boldsymbol{x}_{l,\pi(j)} & l = i, \end{cases} \qquad \text{for} \quad l = 1,\ldots,p, \quad j = 1,\ldots,n. \tag{5.14}$$

In a nutshell, we shuffle the test samples of the predictor variable ξ, generate the forecasts based on the permuted set, calculate the associated CRPS and calculate the difference to the CRPS of the original data. The larger the difference, the more detrimental is the effect of shuffling the feature to the forecast performance, and thus the more important it is.

This procedure can also be applied on a set of $k \leq p$ features Ξ corresponding to the indices $I = \{i_1,\ldots,i_k\}$, which we refer to as *multipass permutation importance* (McGovern et al., 2019). In this case, we do not permute only one feature according to π but instead the entire

set Ξ, i.e.,

$$\tilde{\boldsymbol{x}}^{\pi}_{l,j} := \tilde{\boldsymbol{x}}^{\pi}_{l,j}(\Xi) := \begin{cases} \boldsymbol{x}_{l,j} & l \notin I, \\ \boldsymbol{x}_{l,\pi(j)} & l \in I, \end{cases} \quad \text{for} \quad l = 1, \ldots, p, \quad j = 1, \ldots, n. \tag{5.15}$$

The permutation importance $\Delta(\Xi; \pi)$ is then calculated according to equation (5.13). Last, we calculate the *relative permutation importance* via

$$\Delta_0(\xi; \pi) := \frac{\Delta(\xi; \pi)}{\frac{1}{n}\sum_{j=1}^{n} \mathrm{CRPS}(F(\boldsymbol{x}_{\cdot j}), y_j)}. \tag{5.16}$$

Figure 5.37 shows the relative permutation importance for selected input features and the three NN-based postprocessing methods. There are only minor variations across the three NN approaches, with the wind gust ensemble forecasts providing the most important source of information. To ensure comparability of the three model variants, we here jointly permute the corresponding features of the ensemble predictions of wind gust (mean and standard deviation for DRN and HEN, and the sorted ensemble forecast for BQN). Further results for BQN available in Figure 5.38 indicate that among the ensemble members sorted with respect to the predicted speed, the minimum and maximum value are the most important member predictions, followed by the ones indicating transitions between the groups of subensembles. Again, we find that the standard deviation of the wind gust ensemble forecasts is of no importance for DRN and HEN (not shown).

In addition to the wind gust ensemble predictions, the spatial features form the second most important group of predictors (Figures 5.37 and 5.39). The station ID (via embedding), altitude and stationwise bias are the most relevant spatial features and have a diurnal trend that resembles the mean bias of the EPS forecasts (Figure 5.27), indicating that the spatial information becomes more relevant when the bias in the EPS is larger. Further, the day of the year and the net shortwave radiation flux at the surface provide relevant information that can be connected to the previously discussed evening transition as well as the diurnal cycle. Several temperature variables, in particular, temperature at the ground level and lower levels of the atmosphere, constitute important predictors for different times of day; for example, the ground-level temperature is important for the first few lead times during early morning.

CONCLUSIONS

We have conducted a comprehensive and systematic review and comparison of statistical and ML methods for postprocessing ensemble forecasts of wind gusts. The postprocessing methods can be divided into three groups of approaches of increasing complexity ranging from basic methods using only ensemble forecasts of wind gusts as predictors to benchmark ML methods and NN-based approaches. While all yield calibrated forecasts and are able to correct the systematic errors of the raw ensemble predictions, incorporating information from additional

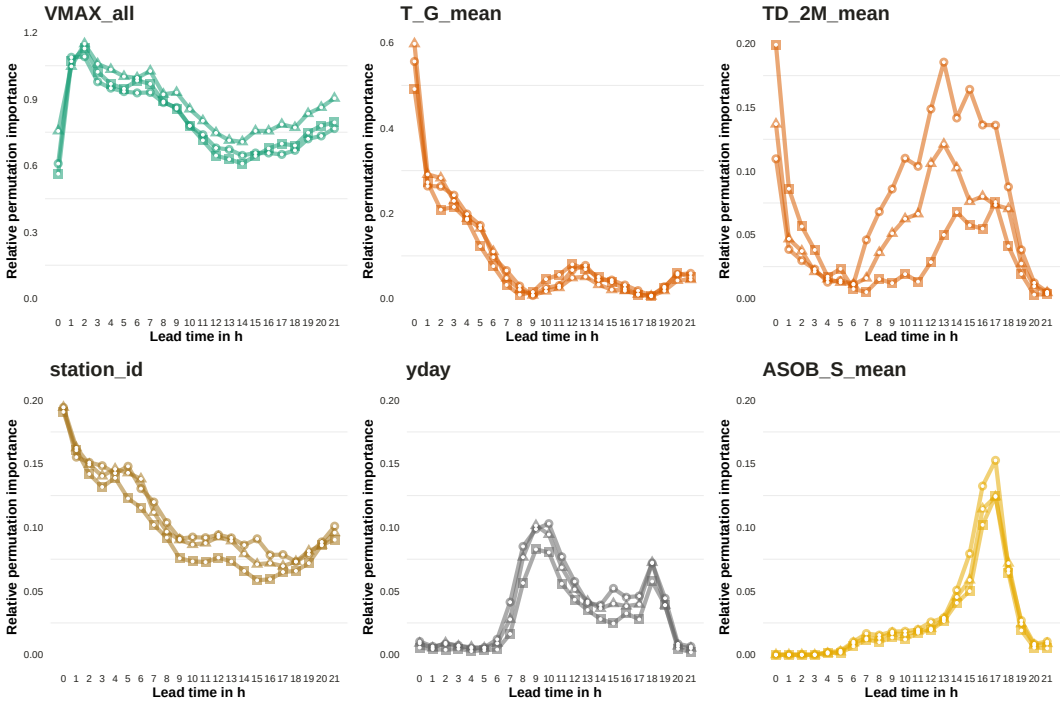

Figure 5.37: Relative permutation importance of selected predictors (Table 5.6) for the three NN-based models dependent on the lead time. Note the different scale of the vertical axes. The abbreviation VMAX_all refers to the multipass permutation of the features derived from the wind gust ensemble. Different symbols indicate the three model variants (∘: DRN, △: BQN, □: HEN).

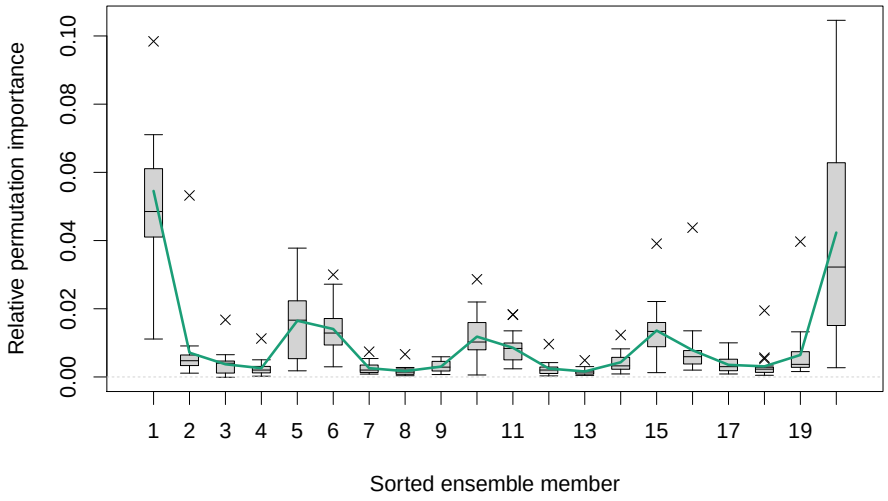

Figure 5.38: Boxplots of the relative permutation importance of the ensemble member predictions of wind gust sorted with respect to the predicted speed for the different lead times in case of BQN. The solid line indicates the mean relative permutation importance.

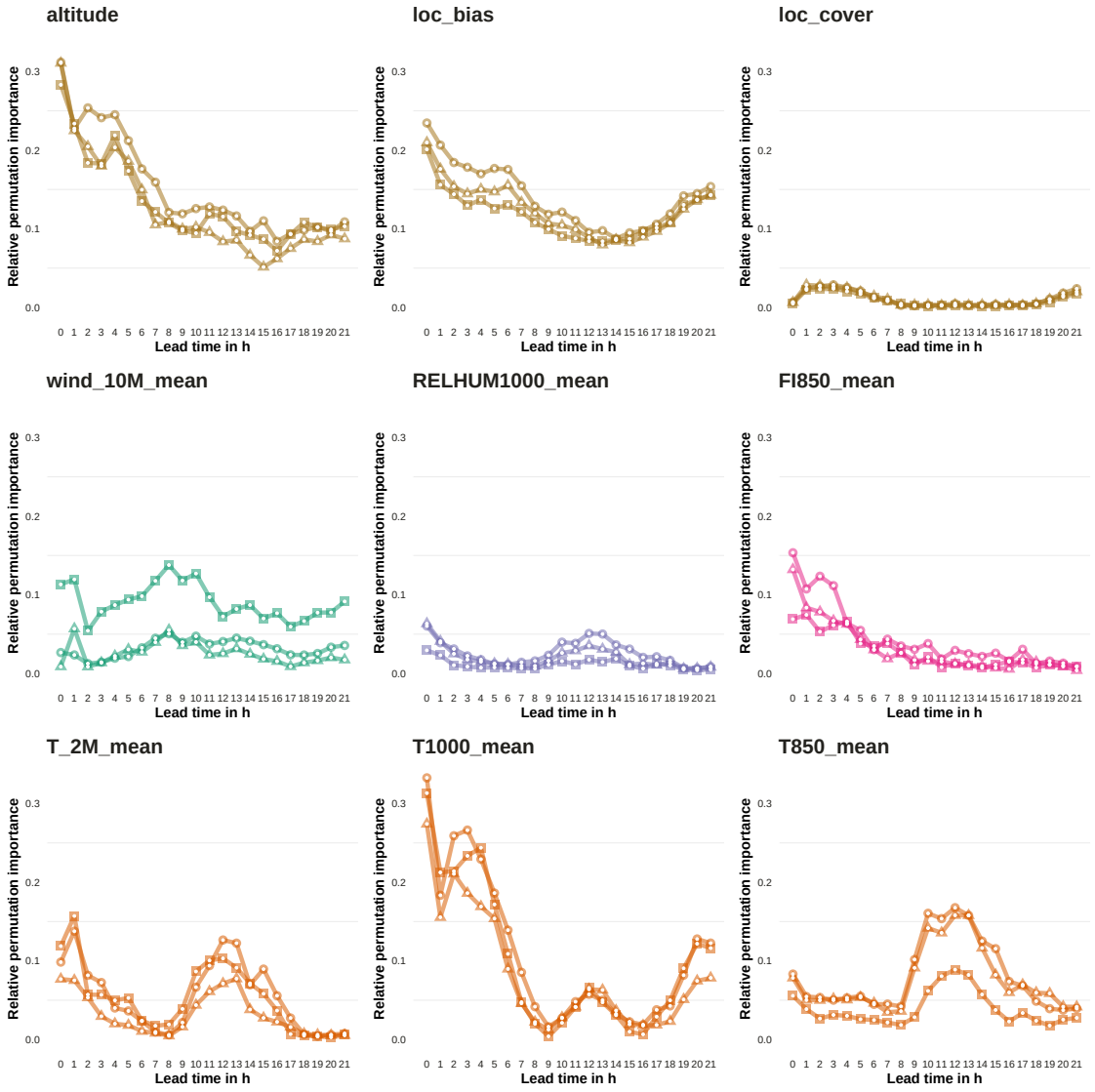

Figure 5.39: Relative permutation importance of selected predictors (Table 5.6) for the three NN-based model variants dependent on the lead time. Note the different scale of the vertical axes. Different symbols indicate the three model variants (○: DRN, △: BQN, □: HEN).

meteorological predictor variables leads to significant improvements in forecast skill. In particular, postprocessing methods based on NNs jointly estimating a single, locally adaptive model at all stations provide the best forecasts and significantly outperform benchmark methods from ML. The analysis of feature importance for the advanced methods illustrates that the ML techniques, in particular the NN approaches, learn physically consistent relations. Overall, our results underpin the conjecture of Rasp and Lerch (2018) who argue that NN-based methods will provide valuable tools for many areas of statistical postprocessing and forecasting.

5.4 Discussion

At the end of the chapter, we synthesize the conclusions from the individual case studies. The first case study exemplifies how a basic statistical postprocessing approach corrects the systematic errors of ensemble forecasts and generates reliable probabilistic forecasts. We implemented an EMOS approach that results in well-calibrated forecasts for various meteorological variables and is used for operational, near real-time postprocessing. In the second case study, we focus on forecasting solar irradiance. Based on two datasets that differ in the target variable, size, spatial domain, temporal resolution, and the underlying NWP model, we adapted the EMOS method for seasonal variations based on different training schemes. Again, the postprocessed forecasts improve the predictive performance with respect to the ensemble forecasts. While the basic approaches in the first two case studies are based only on the ensemble forecasts of the variable of interest, modern ML methods, which allow to incorporate arbitrary predictor variables and to model possibly nonlinear relations to the forecast distribution parameters, are used in the third case study. Comparing the three groups of methods, all methods yield well-calibrated forecasts, but the ML methods improve the performance significantly by using predictor variables beyond the variable of interest. Further, the NN-based postprocessing methods, which outperform the ML approaches established in statistical postprocessing, offer a modular framework that can be tailored to the situation at hand. Here, we implemented a locally adaptive network variant using station embedding and made use of three of different types of forecast distributions.

The postprocessing models for the KIT-Weather portal yield well-calibrated forecasts, but could be developed further by incorporating the forecasts of all of the meteorological variables using EMOS-GB or QRF. Note that this has become feasible only recently as data to a larger extent has been archived. Even further, an NN-based approach making use of the full dataset over location, initialization hours and lead times could be implemented. However, the extension to more complex models includes a trade-off in that a sufficient amount of data is required and that the complexity of the implementation increases. In this operational setting, the incorporation on the portal and the availability in near real-time is a critical requirement.

The datasets used in the solar irradiance case study are somewhat limited in terms of

their temporal extent, in particular the AROME-EPS data. An interesting aspect for future work will be to for example compare different ways of accounting for seasonal variability once longer, ideally multiyear periods of data have become available. As for the near real-time postprocessing, the EMOS models considered could for example be extended using the ML approaches presented in this work. One example of the successful application of NN-based postprocessing that demonstrates the potential value of additional predictor variables is given by Gneiting et al. (2023), who postprocess deterministic NWP forecasts of solar irradiance with the DRN and BQN approach based on the benchmark data from Yang et al. (2022). For similar considerations in the solar irradiance forecasting literature where additional predictors from NWP model output are used, albeit for different types for probabilistic forecasting methods, see, for example, Sperati et al. (2016) and Bakker et al. (2019).

While we did not apply ML-based postprocessing methods in the first two case studies, we found in the third that such advanced method are able to improve the performance significantly, given a comprehensive dataset. But, although we conclude Section 5.3.3 with the statement that NN-based postprocessing methods will provide valuable tools for many areas of statistical postprocessing and forecasting, there does not exist a single best method for most practical applications as all approaches have advantages but also shortcomings. Based on our experiences in the third case study, Figure 5.40 presents a subjective overview of key characteristics of the different methods, ranging from flexibility and forecast quality to complexity and interpretability. We suggest to exploit the full flexibility of NN-based approaches if various additional features and a large training set for model training and validation is available, as it was the case for the wind gust data. If the dataset is more limited, e.g., only given for a small set of stations, the results suggest that QRF and EMOS-GB may still be able to extract valuable information from the additional predictors. However, if only ensemble predictions of the target variable or a small set of training samples are available, more simple and parsimonious methods will likely perform not substantially worse than the advanced ML techniques (see, e.g., the results in Baran and Baran, 2021).

From an operational point of view, one major shortcoming of the postprocessing methods other than MBM is that they do not preserve spatial, temporal or intervariable dependencies in the ensemble forecast. However, in particular in the context of energy forecasting, many practical applications require an accurate modeling of those dependencies (Pinson and Messner, 2018). In terms of the ES, which takes in our case temporal dependencies into account, MBM outperforms EMOS and IDR. However, MBM is still inferior to the ML approaches which do not specifically account for multivariate dependencies, even when focusing on multivariate forecast evaluation. Several studies have investigated approaches that are able to reconstruct the correlation structure (e.g., Schefzik et al., 2013; Lerch et al., 2020) for univariate postprocessing methods. While these techniques require additional steps and therefore increase the complexity of the postprocessing framework, all methods considered here can form basic building blocks for such multivariate approaches. In addition, an interesting avenue for future work is to combine

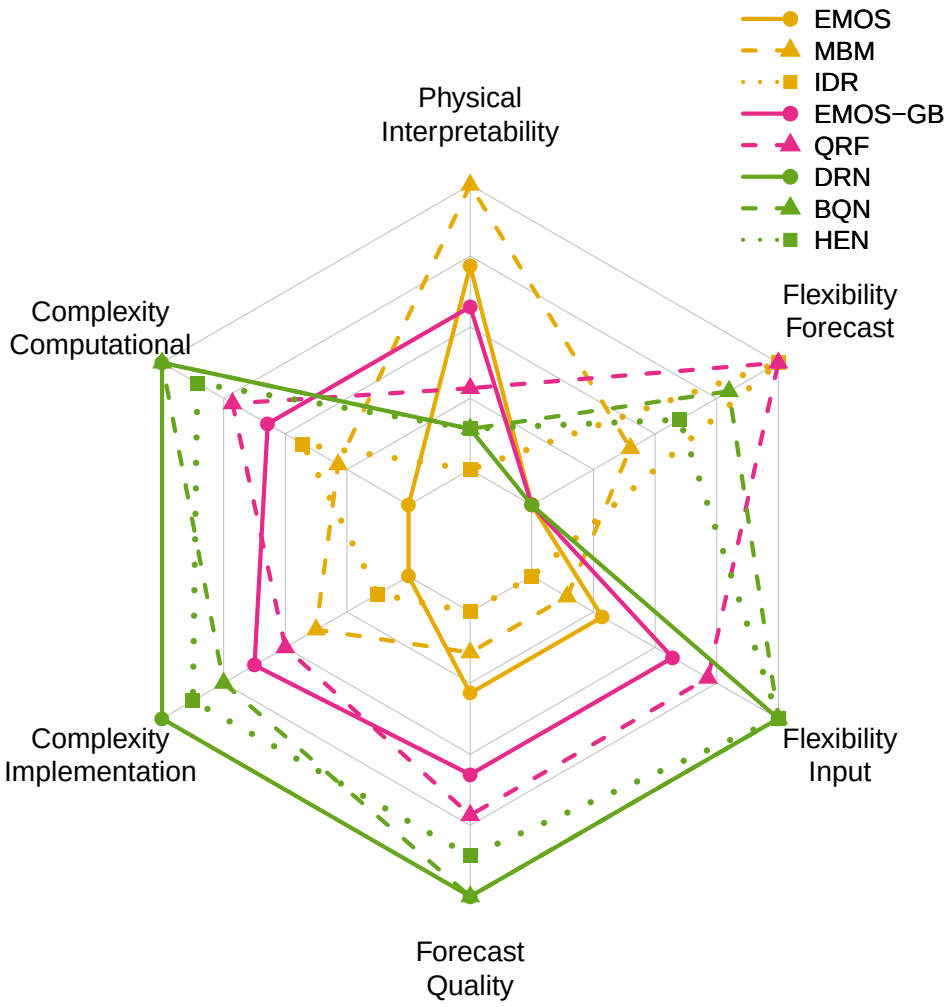

Figure 5.40: Illustration of subjectively ranked key characteristics of the postprocessing methods presented in Chapter 4 in the form of a radar chart. In each displayed dimension, entries closer to the center indicate lower degrees (e.g., of forecast quality). The color scheme distinguishes the three groups of methods, and the different line and point styles indicate different characteristics of the forecast distributions, e.g., solid lines indicate the use of a parametric forecast distribution. Flexibility here refers to the flexibility of the obtained forecast distribution, or the flexibility in terms of inputs that can be incorporated into the model. The component of model complexity is divided into the computational requirements in terms of data and computing resources, and the complexity of the model implementation in terms of available software and the required choices regarding model architecture and tuning parameters.

the MBM approach with NNs, which might allow to efficiently incorporate information from additional predictor variables while preserving the physical characteristics.

A related limitation of the postprocessing methods considered here is that they are not seamless in space and time as they rely on separate models for each lead time, and even each station in case of the basic approaches, as well as EMOS-GB and QRF. In practice, this may lead to physically inconsistent jumps in the forecast trajectories, as described in Section 5.1.3 for the postprocessed precipitation sum forecasts on the KIT-Weather portal. To address this challenge, Keller et al. (2021) propose a global EMOS variant that is able to incorporate predictions from multiple NWP models in addition to spatial and temporal information. For the NN-based framework for postprocessing considered here, an alternative approach to obtain a joint model across all lead times would be to embed the temporal information in a similar manner as the spatial information.

A possible extension of the postprocessing methods presented in this work would be to apply them on the residuals of a linear model instead of the original target variable. This way, the postprocessing methods focus on learning error- rather than scale-dependent relations, thus potentially compensating for the small amount of high wind speeds within the dataset. In particular, QRF would then be able to extrapolate and a more evenly distributed binning scheme for HEN would be obtained. Further, modeling the residuals allows for a better handling of variables differing in their magnitude, such as the solar irradiance variables that are strongly dependent on the diurnal and yearly cycle. For example, Gneiting et al. (2023) use DRN and BQN to postprocess deterministic forecasts for solar irradiance based on the residuals finding that this improves the predictive performance with respect to directly modeling the target variable.

The postprocessing methods based on NNs provide a starting point for flexible extensions in future research. In particular, the rapid developments in the ML literature offer new ways to incorporate additional sources of information into postprocessing models, including spatial information via convolutional NNs (Scheuerer et al., 2020; Veldkamp et al., 2021), or temporal information via recurrent NNs (Gasthaus et al., 2019).

Finally, the development of postprocessing models for solar irradiance was motivated by the aim of improving probabilistic solar energy forecasting. To that end, it would be interesting to investigate the effect of postprocessing NWP ensemble forecasts of solar irradiance for PV power prediction, and for example compare to direct probabilistic models of PV power output (e.g., Alessandrini et al., 2015). In a related study in the context of wind energy, Phipps et al. (2022) find that a two-step strategy of postprocessing both wind and power ensemble forecasts performs best and that the calibration of the power predictions constitutes a crucial step. Ideally, statistical postprocessing of solar irradiance and wind speed forecasts could contribute an important component to modern, fully integrated renewable energy forecasting systems (e.g., Haupt et al., 2020).

CHAPTER 6

Feature-Based Ensemble Postprocessing

A particular challenge for weather prediction is given by the need to better incorporate physical information and constraints into the forecasting models. Physical information about large-scale weather conditions, or weather regimes, forms a particularly relevant example in the context of postprocessing (Rodwell et al., 2018), with recent studies demonstrating benefits of regime-dependent approaches (Allen et al., 2020, 2021). For wind gusts in European winter storms (using the same dataset as in Section 5.3), Pantillon et al. (2018) found that a simple EMOS approach may substantially deteriorate forecast performance of the raw ensemble predictions during specific meteorological conditions, which we will refer to as *high-wind features*.[1] Open research questions include whether similar effects occur also for the more complex NN-based approaches, and how dynamical feature-based postprocessing methods that are better suited to incorporate relevant domain knowledge can be obtained by tailoring the model structure and estimation process.

The final chapter starts in Section 6.1 with an answer to the first question, as we replicate the study of Pantillon et al. (2018) using our NN-based postprocessing methods. The results support the call of Pantillon et al. (2018) for *hybrid* postprocessing, which we vaguely define as approaches that combine a standard model with domain knowledge. To demonstrate the potential of such hybrid models, we will make an excursion to tropical cyclone (TC) forecasting in the North Atlantic, where we will present a model that combines NWP predictions with statistical approaches based on climatological data (Section 6.2). Returning to wind gust prediction, we aim to develop a hybrid postprocessing approach that incorporates information on the relevant weather patterns, here, the high-wind features. As a first step towards such a feature-dependent postprocessing approach, we need to be able to identify the high-wind features. However, this is not straightforward as an expert typically has to identify the features subjectively by hand, e.g., in the case of Pantillon et al. (2018). Hence, an automatic

[1] In this chapter, we will use the term feature to denote only the different high-wind areas, and not as a synonym for the term predictor variable.

procedure that objectively identifies the features is required to develop a feature-dependent postprocessing approach. In Section 6.3, we will present an identification method that combines expertise from ML and meteorology. Based on the identification, a feature-dependent error analysis should assess whether a hybrid approach is actually justified. Here, however, we will skip this part in the interest of brevity and directly focus on experiments towards hybrid models. Therefore, we end the chapter with Section 6.4, where we develop first concepts for feature-dependent postprocessing based on the NN framework introduced in Section 4.3, and assess the predictive performance with a focus on European winter storms.

6.1 Neural network postprocessing within winter storms

Although the postprocessing methods generate accurate and reliable forecasts that outperform the ensemble predictions in Chapter 5, forecast busts have the potential to undermine the acceptance of postprocessing in (operational) practice. Forecast busts, which we vaguely defined in Section 5.1, take different forms, such as for the near real-time postprocessing where one postprocessed forecast of precipitation rate resulted in an extremely large and unrealistic scale parameter. Two other examples in Section 5.1 that can be considered forecast busts are the physically inconsistent precipitation sum or the unrealistically high temperature predictions. In this section, we will focus on cases where postprocessing performs worse than the ensemble and therefore actually degrades predictive performance.

As mentioned in the introduction to this chapter, Pantillon et al. (2018) found that a simple EMOS model deteriorated forecast performance with respect to the ensemble forecasts for some winter storms. In the context of extreme events, which may be subject to an inherent limited predictability, (public) attention focuses on the quality of the associated forecasts. Hence, in these type of situation, forecast busts are especially critical and should be avoided at any cost from the forecasters point of view. We refer to Lerch et al. (2017) for details on the associated challenges they refer to as the forecaster's dilemma.

Here, we will extend the case study on probabilistic wind gust prediction towards the winter storms considered in Pantillon et al. (2018). In the following, we will focus on two of the methods included in the overall comparison of postprocessing methods for wind gusts. Namely, we will apply our EMOS approach to allow comparisons with the original study and to serve as benchmark model for NN-based postprocessing, which outperformed other approaches in Section 5.3. Note that our EMOS approach is not identical to that of Pantillon et al. (2018), and that we will cover the differences over the course of this section. As the performance of the superior NN models DRN and BQN is (almost) indistinguishable within the winter storms, we will consider only DRN in the following, as it is uses the same forecast distribution as EMOS.

Table 6.1: Winter storms selected in Pantillon et al. (2018) together with the selected model initialization and SSI over Germany (names as given by Free University of Berlin; besides Gonzalo, given by the National Hurricane Center). Adapted from Table 1 in Pantillon et al. (2018).

Case	Model initialization			SSI
Joachim	16 December	2011	03 UTC	1.7
Andrea	04 January	2012	18 UTC	2.1
Christian	28 October	2013	00 UTC	5.1
Xaver	05 December	2013	06 UTC	2.9
Gonzalo	21 October	2014	06 UTC	2.5
Elon	09 January	2015	00 UTC	1.9
Felix	10 January	2015	00 UTC	2.6
Niklas	31 March	2015	00 UTC	12.0
Ruzica	08 February	2016	06 UTC	1.0
Susanna	09 February	2016	09 UTC	1.3

6.1.1 Data and model configurations

Our study will be based on the data used in Pantillon et al. (2018), which coincides with that underlying the case study in Section 5.3. Here, we describe how the winter storms were selected and what training and test data is used. For both questions, we follow Pantillon et al. (2018). First, we note that the COSMO model is not only initialized at 00 UTC, the only initialization hour considered in Section 5.3, but instead every 3 hours at 03, 06,..., 21 UTC. The 10 most severe storms in the period from 2011 to 2016 have been selected as case studies, where severity is measured in terms of the *Storm Severity Index* (SSI; Klawa and Ulbrich, 2003), which estimates the impact of a windstorm based on observed surface gusts, here, in meteorological terms.[2] For each of these storm, Pantillon et al. (2018) pick one initialization time and consider forecasts for all available lead times from 0 to 21 hours, where the initialization time was chosen such that the highest intensity is reached after lead times from 12 to 15 hours. Table 6.1 provides an overview of the selected winter storms.

In the overall comparison, we used the period from 2010 to 2015 for training, including the validation year 2015, and 2016 for testing. As the selection of winter storms covers the entire period, we cannot use the models from the case study for prediction and need to find another approach to split the data. To do so, we use a yearly CV approach where we hold out the year of occurrence for each storm and use the remaining years for training. Further, the year 2016 is used as validation period, besides for storms in that year, where we use 2015 for validation. Hence, for storms in 2016, the underlying data partition is identical to that in Section 5.3. As in previous applications, a separate model is estimated for each lead time. In contrast to Section 5.3, the initialization times chosen for the selected storms are based on

[2]Originally, the SSI was developed to estimate the damage to buildings and infrastructure, also taking into account the population density as indicator for insured values.

different hours of the days (Table 6.1). Following the discussion in Sections 5.1 and 5.2, we treat each initialization hour separately and train a separate model. Hence, storms in different years and/or based on different initialization hours are not predicted using the same model instances. However, when two storms coincide in both categories, the same model is used, as in case of Elon, Felix and Niklas.

In order to ensure comparability to Section 5.3, we do not modify the two postprocessing methods and apply them directly as before. This includes the fact that we do not tune the hyperparameters separately but instead use the configurations described previously. However, as the data partition scheme is consistent with Section 5.3, we do not expect a negative effect on the predictive performance. As mentioned before, the EMOS approach used in Pantillon et al. (2018), which we refer to as EMOS-0, differs from our approach used in Section 5.3. The main difference is that Pantillon et al. (2018) use a rolling 30-day window for training of EMOS-0, while our approach is based on a seasonal training period over five years.

6.1.2 Results

When comparing the evaluation metrics in Figure 6.1 with Pantillon et al. (2018, Figure 5), we find that the results of the two EMOS variants are consistent. The CRPS, CRPSS, forecast bias and PI length/forecast uncertainty are of the same order and behave largely identical.[3] Both EMOS approaches are subject to the forecast bust for Christian, which comes with a large negative bias, and Andrea is the only storm that results in a positive bias for both. Further, the two approaches result in a negative skill for Felix. However, there are also differences. Only for EMOS-0, we observe a forecast bust in case of Xavier, and, only for our EMOS approach, we observe a forecast bust in case of Niklas, which is even larger than for Christian. Despite the minor differences, we conclude that the results are coherent and that changes are a result of the different training schemes.

Next, we compare the performance of EMOS and DRN in the winter storms with that over the entire year. Note that, in contrast to the winter storms, the general comparison in Section 5.3 is based only on initialization times at 00 UTC, which reduces comparability over lead times due to different effects of the diurnal cycle, and that the sample size is much smaller here. Figure 6.2 shows four of the evaluation metrics we also considered in Figure 5.30, averaged over all winter storms. As expected, when looking at extreme events, the forecast uncertainty increases and larger scores are observed. Both the CRPS and CRPSS are not at the level of the general comparison, and an increase in PI length reduces the sharpness. Further, we observe the spin-up effect in a sudden jump in the bias and PI length of the ensemble forecasts for the 1 hour forecast.

Regarding the differences between EMOS and DRN, we find a similar gap of around 10% in

[3] Pantillon et al. (2018) do not evaluate the forecast uncertainty based on the PI length but instead the standard deviation (referred to as spread). Hence, Figure 3d in Pantillon et al. (2018) and the PI length of the EMOS approach in Figure 6.1 are on different scales. Still, they behave similarly.

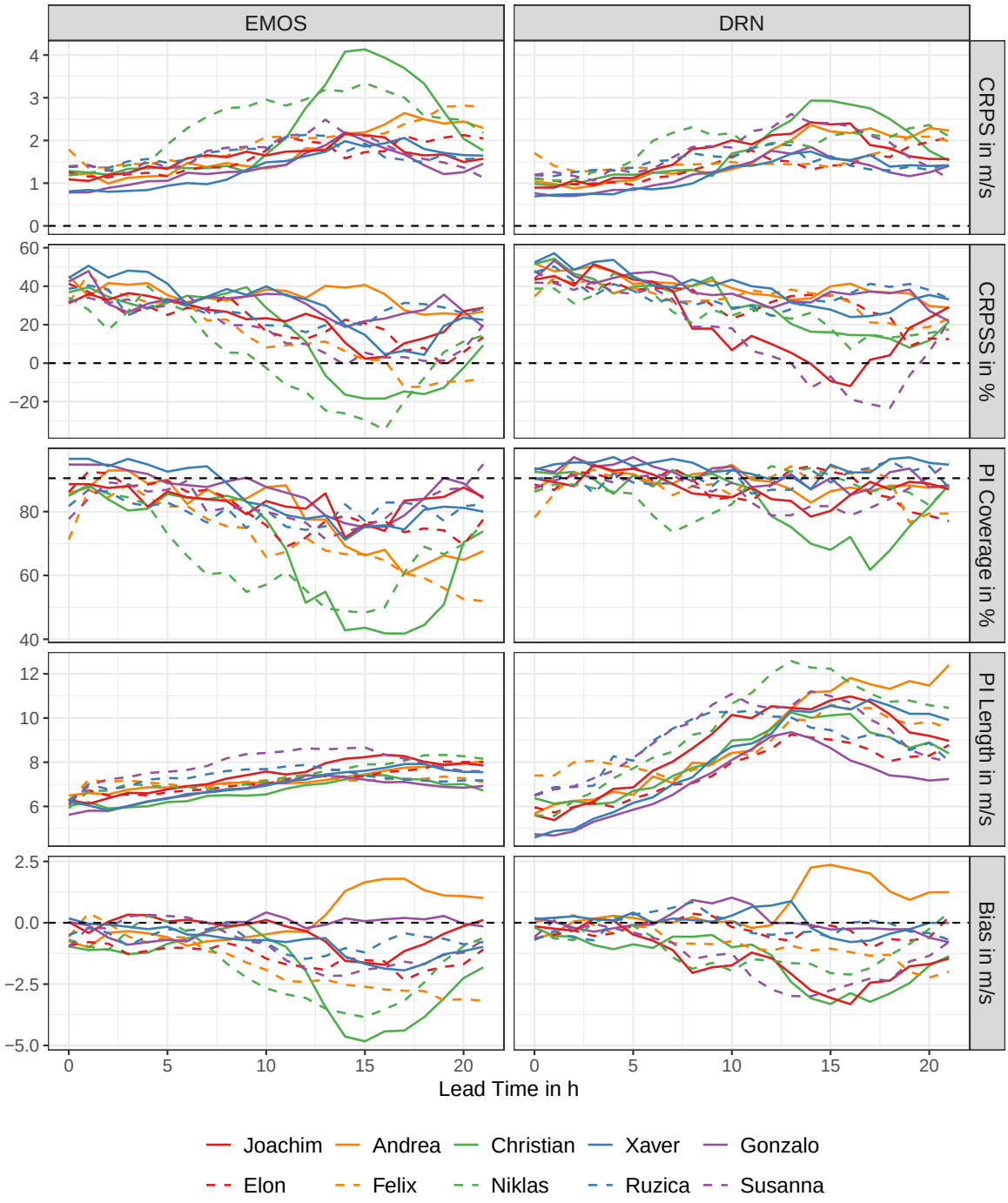

Figure 6.1: Mean CRPS, CRPSS with respect to the raw ensemble predictions, mean PI coverage, mean PI length and mean bias of EMOS and DRN as functions of the lead time, averaged over all stations for each winter storm.

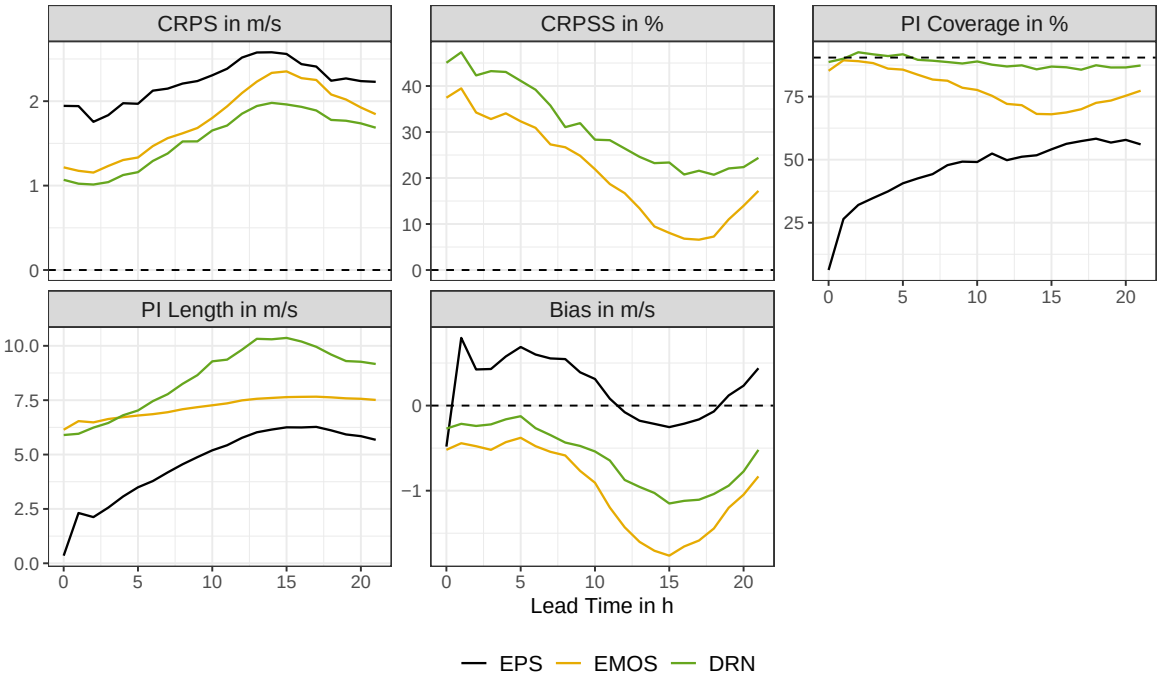

Figure 6.2: Mean CRPS, CRPSS with respect to the raw ensemble predictions, mean PI coverage, mean PI length and mean bias of EMOS and DRN as functions of the lead time, averaged over all winter storms and stations.

the CRPSS, which is largest when the storms are most intense. As in the general comparison, both methods predict smaller median wind speeds than the ensemble forecasts. While this reduction resulted in (nearly) unbiased forecasts in the general comparison, the methods strongly underforecast the wind gusts in the winter storms. Comparing EMOS and DRN, we find that the bias of the EMOS forecasts is larger, especially when the intensity peaks. When the storms are most intense, the DRN forecasts are still reasonably well-calibrated, as the PI coverage stays close to the nominal value in contrast to EMOS. In addition to a smaller bias, wider PIs are a reason for the higher PI coverage of the DRN forecasts. In the general case, DRN provides on average sharper forecasts than EMOS. Here, however, EMOS generates smaller PIs than DRN. But comparing the behavior of the forecast uncertainty with that of the ensemble predictions, we find that while DRN resembles the uncertainty of the ensemble predictions, EMOS adapts only slightly. Hence, we conclude that DRN describe the forecast uncertainty much more adequately.

While the postprocessing methods perform well averaged over all cases, a separate investigation of the individual winter storms is crucial due to the diversity of the selected cases. Figure 6.1 shows the evaluation metrics for each storm separately, both in case of EMOS and DRN. In addition, Figures 6.3 and 6.4 simultaneously compare the two methods with the ensemble forecasts for each storm separately. Note that at most 175 forecasts are available for each lead time, depending on the number of stations with missing values.

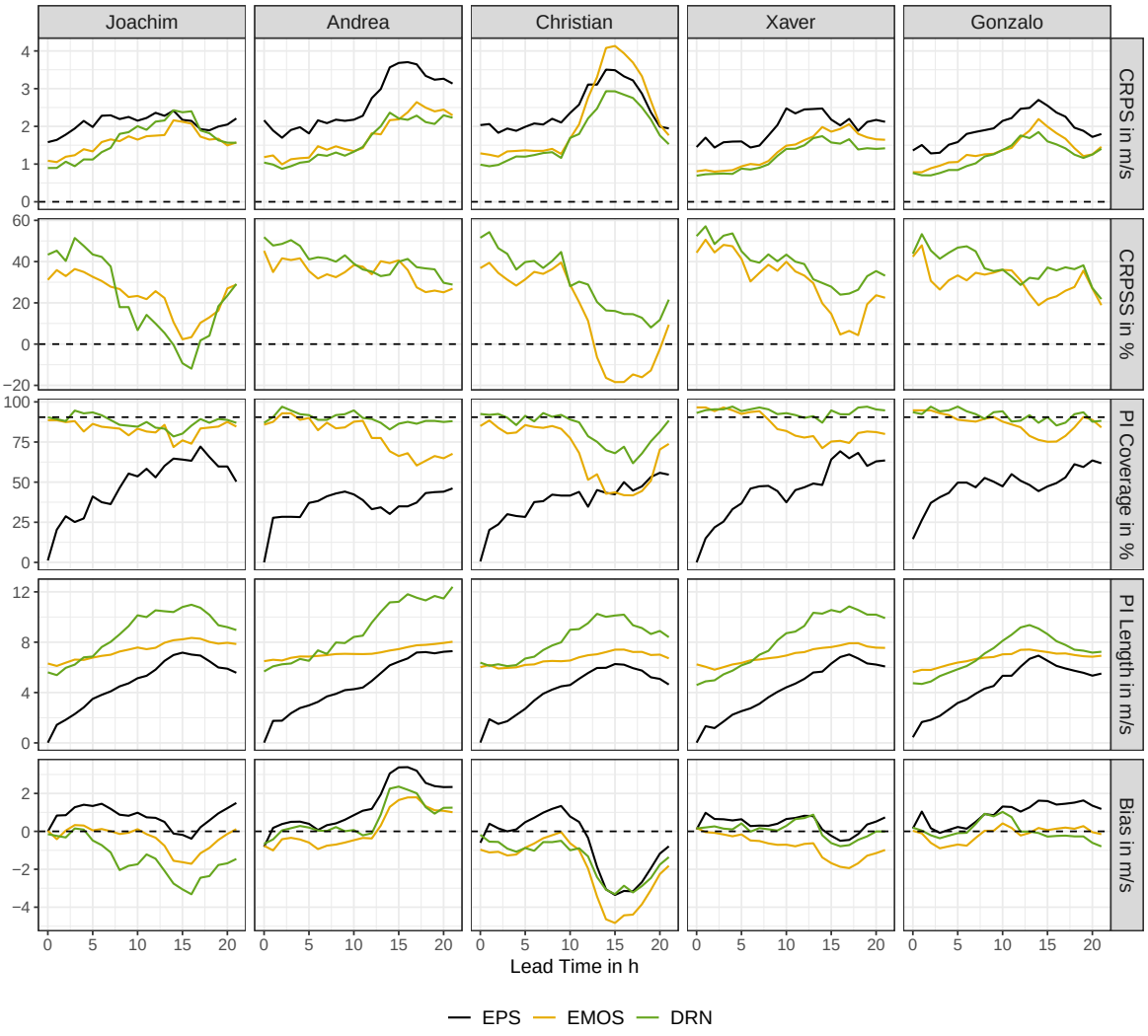

Figure 6.3: Mean CRPS, CRPSS with respect to the raw ensemble predictions, mean PI coverage, mean PI length and mean bias of EMOS, DRN and the raw ensemble predictions as functions of the lead time, averaged over all stations plotted separately for each winter storm within the period from 2011 to 2014.

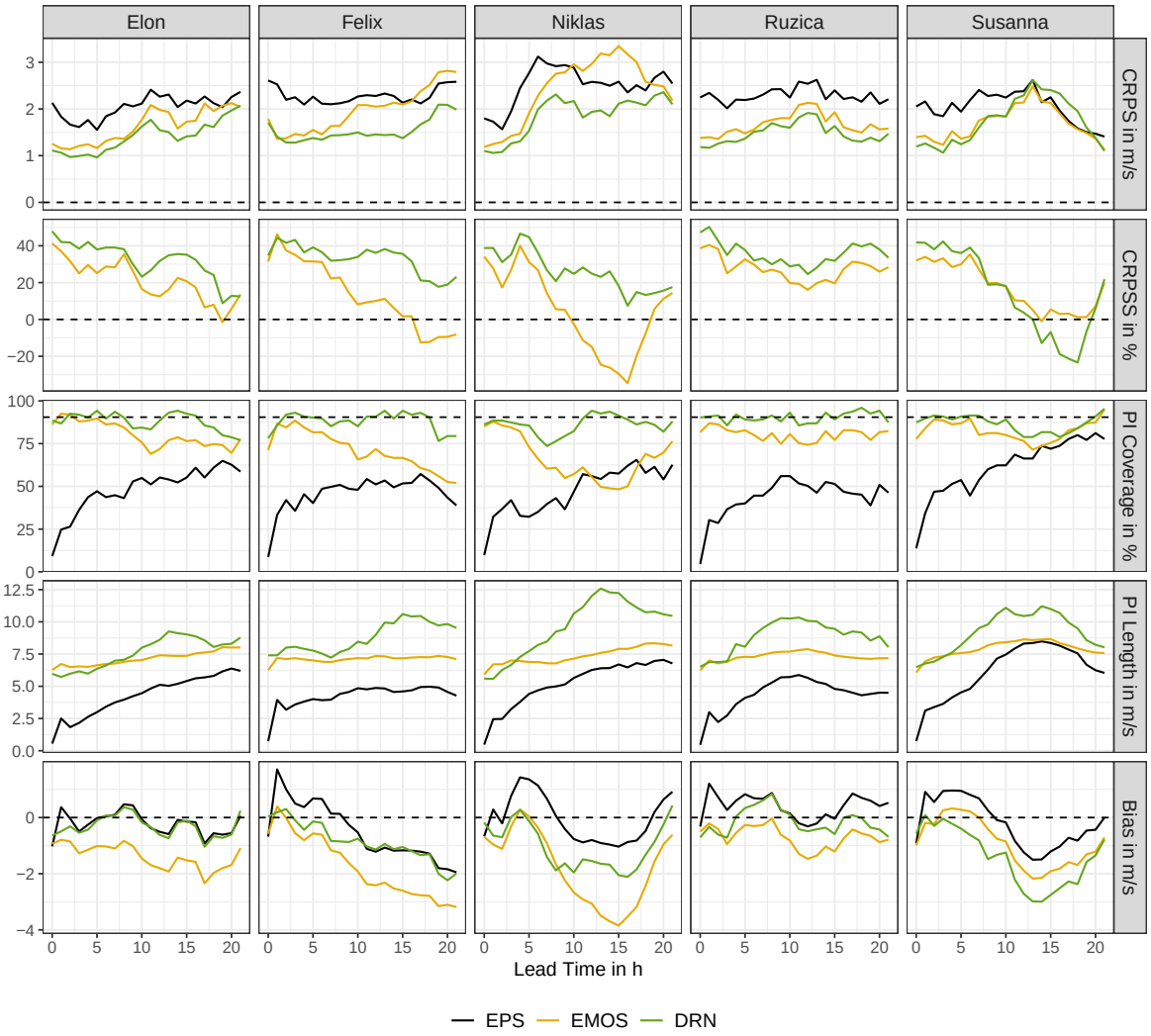

Figure 6.4: Mean CRPS, CRPSS with respect to the raw ensemble predictions, mean PI coverage, mean PI length and mean bias of EMOS, DRN and the raw ensemble predictions as functions of the lead time, averaged over all stations plotted separately for each winter storm within the period from 2015 to 2016.

As mentioned in the comparison of EMOS and EMOS-0, there are three forecast busts for EMOS (Christian, Niklas, Felix). In contrast, two are observed for DRN (Joachim, Susanna). Interestingly, the underlying storm cases differ. Still, the magnitude of the forecast busts is smaller in case of DRN. When comparing the CRPS of DRN with that of the ensemble for the individual storms (Figures 6.3 and 6.4), we find that, for all cases but Susanna and Joachim, a constant gap in the CRPS is present. On the other side, the CRPSS of EMOS becomes worse as the storm intensifies. When comparing the PI lengths for the individual storms, we find the same pattern as for the averages. The DRN forecasts resemble the forecast uncertainty of the ensemble, while the EMOS forecasts adapt only slowly. The behavior of EMOS coincides with a decrease in the CRPS, hence the method fails to adapt to the arising forecast uncertainty and does not increase the PI length sufficiently. In case of the forecast bias, we find that the reduction of the median forecast performed by EMOS is consistent over all lead times with a constant gap for almost all cases. In contrast to EMOS, the correction of the median forecast changes in case of DRN for each storm and over lead time. Overall, the predicted wind speed is still corrected downwards (as seen in Figure 6.2). Note that the varying bias corrections of DRN do not originate from the underlying model instances, as we observe this behavior also in case of Elon, Felix and Niklas, which are based on the same network models.

In the following, the individual storms are investigated in more detail, where we focus on DRN:

- **Joachim, Susanna:** In these two cases, the skill of DRN vanishes and becomes negative. For both, DRN strongly underforecasts the observed wind gusts as a result of a downward correction of the ensemble median forecast, which already has a negative or no bias for the late lead times. Interestingly, these cases are the only two where DRN applies a larger bias correction than EMOS. Although EMOS is not subject to a forecast bust for these two storms, almost no skill is achieved for the later lead times.

- **Andrea:** Andrea is the only storm, where a positive bias is observed for EMOS and DRN. Still, both postprocessing methods reduce the bias with respect to the ensemble median, which overforecasts the observed wind speeds strongly. As long as DRN is not more biased than EMOS, it performs better due to a more adequate forecast uncertainty.

- **Xaver, Elon, Ruzica:** For these three storms, DRN outperforms EMOS consistently over all lead times and applies only a small bias correction with respect to the ensemble forecasts. EMOS consistently underforecasts the observed wind speeds.

- **Christian, Gonzalo, Niklas:** Here, we can observe an interesting pattern of DRN, that is, the bias correction changes when the intensity of the storm increases. In all cases, this reduction improves the predictions as it reduces the absolute bias compared to the (hypothetical) application of an unchanged bias correction with respect to the lead time. In case of Gonzalo, the predicted median speed is corrected downwards. As

the ensemble overforecasts the observed wind speed, this reduces the bias and improves the predictive performance. For Christian and Niklas, the two forecast busts of EMOS, DRN corrects the predicted wind speeds in the opposite direction, that is, upwards, and thus prevents a forecast bust.

– **Felix, Niklas:** In both cases, we observe that the PI length of DRN increases at late lead times, although the ensemble range does not increase (at the same rate). Thus, DRN identifies that the forecast uncertainty should be increased more than suggested by the ensemble predictions.

Concerning the individual winter storms, we conclude that the corrections performed by EMOS are consistent over lead time and storm cases. This is not surprising, as the EMOS model is based only on the wind gust ensemble predictions via a linear link function and does not model nonlinear, flexible relations in contrast to the NN-based postprocessing methods. DRN seems to be able to identify cases, in which different bias and also dispersion corrections are required for the ensemble predictions. However, this does not always work, as we still observe two forecast busts in case of Joachim and Susanna.

To better understand the corrections applied by DRN, we investigated the importance and effect of the predictor variables using methods for interpretable ML (Molnar, 2018).[4] However, these experiments remained inconclusive, as we found no systematic effects of the predictor variables explaining the behavior of the DRN models. One aspect that hindered inference was that we estimate separate models for (almost) each winter storm and lead time, that are themselves based on deep ensembles of 10 individual networks.

Overall, we conclude that NN-based postprocessing outperforms the basic EMOS approach also within the winter storms. In contrast to EMOS, DRN is well-calibrated when the storm intensifies, as the PI coverage remains close to the nominal level. Further, DRN quantifies the forecast uncertainty adequately, again, in contrast to EMOS. Still, not only EMOS but also DRN is subject to forecast busts, where postprocessing performs worse than the ensemble predictions. In their study, Pantillon et al. (2018) investigated the selected winter storms with respect to the meteorological conditions and associated wind gust-generating mechanisms involved, and conjectured that these high-wind features are associated with the forecast busts of their EMOS approach. Therefore, they called for feature-dependent postprocessing that accounts for the high-wind features. But before we follow this pathway by developing an identification method in Section 6.3, we will demonstrate the potential of building hybrid forecasting models.

[4] In particular, we applied the permutation importance technique used in Section 5.3 as well as partial dependency plots, a technique that will be introduced at a later point of this chapter in Section 6.3.

6.2 Excursion: Hybrid forecasting of tropical cyclones

At this point, we want to highlight a study on probabilistic prediction of TC occurrence in the North Atlantic, which demonstrates the potential of combining different sources of information in a hybrid model. Despite the use of methods from statistical learning to improve upon NWP model forecasts, we refer to this section as an excursion, since we do not stay in the postprocessing setting of Chapter 4, as we consider a binary instead of a continuous target variable. By systematically comparing forecasting methods that differ in the inherent type of information, we want to motivate the usage of hybrid models to improve predictive performance. In the following, we shortly describe the setting, data and types of models used, including benchmark and statistical models. Then, results of the systematic model comparison are presented, followed by a summary and conclusive remarks. As this section is intended to provide only a glimpse of the entire underlying study, we will keep the descriptions short and refer to Maier-Gerber et al. (2021) for details.[5]

For decades, there has been a parallel development of predictions for individual TCs made by operational forecast centers for lead times of a few days on the one hand, and seasonal predictions of integrated TC activity on the other.[6] This coexistence is due to the *subseasonal predictability gap* (Vitart et al., 2012; Robertson et al., 2020), which has raised broad attention and efforts to bridge only in recent years, where we refer to subseasonal predictions as forecasts with lead times from two to five weeks ahead. Because of the lack of skillful models, potential sources for subseasonal predictability of TC activity have become a research focus. Nowadays, NWP models are often integrated to subseasonal or seasonal forecast horizons, and have been systematically evaluated in terms of predictive skill for different TC occurrence measures in several studies (Lee et al., 2018, 2020; Gregory et al., 2019). Lee et al. (2018) found that the *subseasonal-to-seasonal* (S2S; Vitart et al., 2017) models generally have little to zero skill in predicting TC occurrence from week two on relative to climatological forecasts. For the North Atlantic, which is the domain considered in this section, they stated that actual and potential model skills are very close, suggesting that hardly any improvement can be achieved with current NWP models.[7]

Inspired by the example of numerous statistical forecast models for seasonal forecasting, Leroy and Wheeler (2008) followed a different approach and developed logistic regression models based on past data to produce probabilistic forecasts of weekly TC genesis and occurrence up to seven weeks in advance. Comparing against ECMWF model predictions,

[5] Note that the publication underlying this section is a collaboration with meteorologist (see Section 1.1 for my contributions).

[6] The term integrated TC activity refers to both the number and intensity of TCs, integrated over the prediction horizon.

[7] Lee et al. (2018) measure potential model skill based on the approach of Buizza (1997), who investigates a model-dependent limit of forecast skill. Note that this does not correspond to the potential predictability of the underlying target variable, here, TC activity.

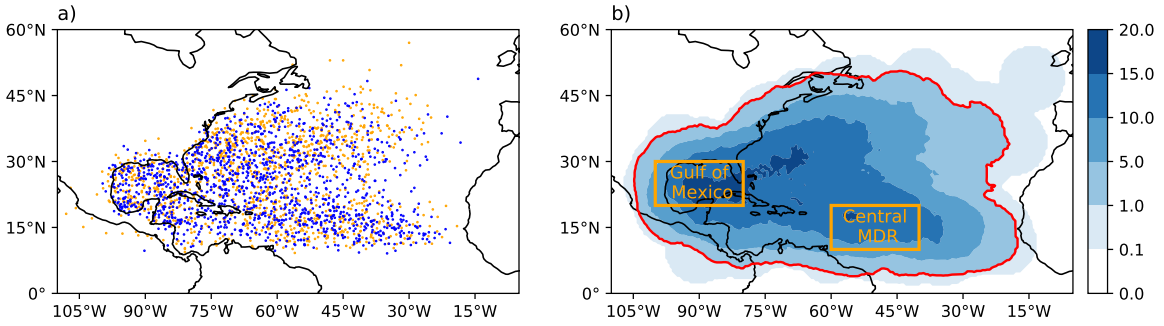

Figure 6.5: (a) 1968–1997 (orange dots) and 1998–2017 (blue dots) IBTrACS tropical cyclone positions at 00 UTC during the North Atlantic hurricane season (June–November) for intensities of at least tropical storm strength. (b) Relative frequency of TC occurrence (%) based on the definition in the data paragraph. Note that interval boundaries are not equidistant. The red contour highlights the area where TCs occur at a rate of more than 1%. Orange boxes enclose the subregions used for model validation.

Vitart et al. (2010) identified the statistical approach from Leroy and Wheeler (2008) to perform better from week two on.

Even though Lee et al. (2018) concluded that the (dynamical) S2S models lack skill to forecast North Atlantic subseasonal TC genesis, these models may be able to predict subseasonal environmental conditions favorable for TC genesis to a sufficient degree, so that predictors can be generated and fed into statistical models. Such a statistical-dynamical (or hybrid) forecast model is thought to combine the strengths of each individual model, and thus to increase model skill. Here, we present a hybrid model, using a variety of predictors known to precondition and modulate environments that are prone to TC occurrence.

DATA: TC OCCURRENCE

The basis for deriving the TC occurrence is the *International Best Track Archive for Climate Stewardship* (IBTrACS; Knapp et al., 2010, 2018) dataset version 4. To account for TC occurrence, cyclones are required to be tropical in nature and to exceed at least tropical storm strength ($\geq 34\,\text{kt}$; $1\,\text{kt} \approx 0.51\,\text{m/s}$). Although the IBTrACS dataset comes with a 3-hourly temporal resolution, only 00 UTC instances of cyclone track positions are taken into account to allow for a systematic comparison with the lowest temporally resolved benchmark model, the S2S TC tracks (see benchmark models). Figure 6.5a shows the North Atlantic cyclone positions, that fulfill the stated criteria for the periods used for model validation, training of the statistical models, and for generating the climatological models, respectively.

To take account of the reduced predictability on subseasonal timescales, TC occurrence (hereafter alternatively referred to as 'target variable') is created by means of a coarser spatio-temporal evaluation, which is over periods of one week and within a certain spatial

area. For a given forecast week, a grid point is considered to feature TC occurrence, if at least one TC occurs within a radial distance of 7.5°. Based on the dichotomous target variable, Figure 6.5b presents a map of the resulting relative frequencies of TC occurrence, which can also be interpreted as an density plot of the occurrence.

Data: Predictor variables for the statistical models

The difference between the statistical-dynamical approach and the purely statistical approach developed is merely in the underlying data, from which predictors are generated. The purely statistical models are trained on ERA5 (Hersbach et al., 2020) data, whereas predictors for the statistical-dynamical models are generated from S2S ECMWF ensemble reforecasts.[8]

The S2S reforecasts are produced twice per week (Mondays and Thursdays) with one control plus 10 perturbed forecasts, ranging out to 46 days. Originally calculated with a horizontal grid spacing of 16 km for the first 15 days and 31 km afterwards, S2S model output is archived with daily values at 00 UTC on a regular $1.5° \times 1.5°$ grid, which is considerably coarser compared to ERA5. For the sake of consistency, both datasets are therefore used with this coarser grid spacing and temporal resolution. To ensure that the S2S-based predictors are not subject to biases, a mean bias correction was applied to all variables, from which predictors were directly generated.

Benchmark models

As demonstrated in Chapter 5, an integral part of model development is to compare a newly generated model with those that are well-established and/or different in their approach. To justify the application of a new model, it should perform better than the benchmark. With climatological and NWP models, two distinct types of benchmark models are employed in the following to put into relation the performance of the statistical models developed.

Climatological models are used as the first type of benchmark to allow for a comparison with predictions based on long-term statistics of TC occurrence, i.e., on its climatology. Because those statistics are calculated over a set of past realizations drawn from the underlying distribution of the target variable, climatological forecasts are inherently independent of the current state of the atmosphere. Moreover, they are unbiased if trends and/or regime changes are negligible. If so, there are no restrictions regarding lead time, and forecasts are thus independent of forecast week. The climatological models used here are derived from the IBTrACS dataset for the period 1968–2017. The simplest approach to generate a climatological statistic is to average TC occurrence over the 50 North Atlantic hurricane seasons considered. This approach yields a *mean seasonal climatology* (MSC), where constant forecasts are predicted throughout the season. A more adaptive strategy to take into account

[8] ERA5 is short for ECMWF Reanalysis version 5. While reanalysis data describes the past state of the atmosphere as close as possible using all information available at the present day, reforecasts replicate the forecasts the current model would have generated in the past.

seasonal variations is to average over years for every day of year separately, resulting in a *climatological seasonal cycle* (CSC). Seasonal fluctuations indicate that the 50-year period is not sufficient to generate a robust climatology, since one would expect the observed relative frequency to not vary much for neighboring days in the year. To mitigate the adverse effect of too small sample sizes, a smoother and more representative CSC (hereafter referred to as *CSCopt*) was constructed by applying a moving average.

To compare with predictions directly obtained from a state-of-the-art NWP model, a second type of benchmark is created by calculating probabilities for TC occurrence from all 00 UTC instances of the TC tracks identified in the S2S ECMWF 1998–2017 ensemble reforecasts (hereafter referred to as *S2STC*). Each ensemble member either predicts the occurrence of a TC or not, therefore we can derive a probability forecast by averaging over the 11 ensemble members. Note that the probability forecasts take the values $0/11, 1/11, \ldots, 11/11$.

Just like the short- and medium-range NWP predictions considered in Chapter 5, S2STC forecasts are frequently not calibrated. Therefore, we have tested different statistical post-processing techniques to correct for potential miscalibration. Note that TC occurrence is a dichotomous target variable, and not (discrete-)continuous as those in Chapter 5. Hence, we cannot apply the methods presented in Chapter 4, but instead we consider techniques tailored to calibration of probability forecasts, which have briefly been addressed in Section 3.5, such as the beta-transformed LP (Ranjan and Gneiting, 2010), beta calibration (Kull et al., 2017) or logistic calibration (Platt, 1999), but also IDR using the probability forecast as sole predictor. For this purpose, IDR turned out to perform best. In the application on probability forecasts, IDR is equivalent to *isotonic regression*, a common approach for calibration of probabilities in the ML literature (e.g., Guo et al., 2017). Based on the natural assumption that a higher forecast probability is associated with a higher event frequency, IDR here learns a step-function that is used to transform the S2STC forecasts to calibrated probability forecasts.[9] To increase robustness, forecasts from all grid points of a given validation subregion are pooled for training of IDR, hence, a global training approach.

Statistical model development

If the target variable is dichotomous, logistic regression models are commonly trained to map linear combinations of continuous predictor variables to a probability via the logit function (Hastie et al., 2009).[10] The logistic regression model can be formulated as

$$\pi\left(\boldsymbol{x};\beta_0,\boldsymbol{\beta}\right) \;=\; \text{logit}^{-1}\left(\beta_0+\boldsymbol{x}^T\boldsymbol{\beta}\right)=\left(1+\exp\left(-\beta_0-\boldsymbol{x}^T\boldsymbol{\beta}\right)\right)^{-1}, \tag{6.1}$$

[9] In this particular application, the probability forecasts only take 12 values. Hence, in essence, we only need to find calibrated probabilities for each of the 12 predictions derived from the ensemble. In case of IDR, these values simply correspond to the average occurrences observed in the training data, under the constraint of monotonicity.

[10] In Section 5.1, we use a multivariate logistic regression variant for postprocessing of cloud cover (equation (5.2)).

where π denotes the estimated probability of the target variable being 1, $\boldsymbol{x} \in \mathbb{R}^p$ the vector of the predictor variables, $\beta_0 \in \mathbb{R}$ the intercept, and $\boldsymbol{\beta} \in \mathbb{R}^p$ the vector including the regression coefficients of the predictors. Using the LIBLINEAR solver (Fan et al., 2008), we estimate the coefficients by solving the optimization problem

$$\min_{\beta_0, \boldsymbol{\beta}} \quad \frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{\beta} + \sum_{i=1}^{n} \mathrm{LogS}\left(\pi\left(\boldsymbol{x}_i; \beta_0, \boldsymbol{\beta}\right), y_i\right), \tag{6.2}$$

where $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ is the training data. The second term corresponds to MLE (and optimum score estimation), while the first term corresponds to a l_2-penalty, which keeps the coefficients of the predictors small and thus prevents the model from overfitting. The minimization is stopped, if either the difference between the losses of two consecutive iterations drops below a tolerance of 10^{-4}, or a maximum number of 100 iterations is reached. To support faster convergence of solutions for model coefficients, predictors are standardized on the respective training set.

Training a logistic regression model on the full variety of predictors available does not necessarily lead to the best predictive performance. Optimal predictor subsets for the statistical-dynamical and purely statistical approach, respectively, are therefore determined using a sequential forward predictor selection (e.g., Hastie et al., 2009). This selection process is conducted separately for the two subregions considered (Figure 6.5), namely, the Gulf of Mexico and the central main development region (MDR), and grid points are pooled for each subregion to make selections more robust. To guarantee that the logistic regression models do not perform worse than the climatological benchmark models, a CSCopt predictor is kept fixed, a priori. This initial minimal subset is then extended by the one predictor that minimizes the average *Akaike information criterion* (AIC; Akaike, 1974; Hastie et al., 2009) of a 5-fold CV on the training period. For a logistic regression model with p predictors, the AIC is defined as

$$\mathrm{AIC} \quad = \quad 2\frac{1}{n}\sum_{i=1}^{n} \mathrm{LogS}\left(\pi\left(\boldsymbol{x}_i; \beta_0, \boldsymbol{\beta}\right), y_i\right) + 2\frac{p}{n}, \tag{6.3}$$

where $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ is a set of n forecasts and observations. We chose AIC as our scoring metric since it reduces overfitting by penalizing larger numbers of predictors, in addition to the term for the model's performance. The extension of the subset is repeated until all candidate predictors are integrated. Then, the optimal subset of predictors is finally identified by the lowest AIC achieved. This forward selection is preferred over a backward selection (i.e., successively removing predictors) to keep the number of optimal predictors as small as possible but as large as necessary. Hence, 20 predictor subsets are obtained that are found to be highly consistent, being in complete agreement for the central MDR, and differing in only one predictor at week two for the Gulf of Mexico.

Validation strategy and training of the statistical models

A systematic comparison of the different model approaches requires a common strategy for validation. While forecasts from the climatological benchmark models can be issued every day, the NWP-based benchmark model and the predictors for the logistic regression models rely on the twice-weekly run and disseminated S2S ECMWF forecasts, thus, posing the stronger limitation to a validation dataset. Starting from each of these S2S reforecast initialization dates, for every model, forecasts are generated for the first five consecutive weeks, i.e., days 0–6, 7–13, ..., 28–34. However, forecasts are only considered for validation if the middle of the respective forecast week falls into the North Atlantic hurricane season, which runs from 1 June to 30 November. This yields a total of 1,040 (52 reforecasts per season $\times$ 20 seasons) validation instances, for which S2S ECMWF reforecasts are available. In contrast to the NWP-based benchmark models, the climatological and logistic regression models require a training dataset that is independent of the validation dataset. To fully exploit the relatively small number of S2S reforecasts for both purposes, a 20-fold CV is applied, so that every season can be successively validated, while the statistical models are being trained on the remaining 19 seasons. To avoid training statistical models with too great imbalances between TC occurrence and nonoccurrence in the target variable, a fraction of at least 1% is required to feature TC occurrence, which is the case for the Gulf of Mexico and central MDR subregions. Because the climatological models (and thus the base predictor) share the underlying dataset with the target variable, the CV strategy necessitates the climatologies to be calculated separately for every fold, leaving out the data of the season to be forecast and validated. Although a separate statistical model is trained for every grid point and target forecast week, the generated forecasts are pooled for each of the two subregions, to allow for more solid conclusions during the validation discussed in the latter.

Model comparison

Before all models are validated in terms of skill, calibration and sharpness of the probability forecasts are analyzed qualitatively via CORP reliability diagrams (Section 2.1). Figure 6.6 shows CORP reliability diagrams for the Gulf of Mexico and central MDR week-four forecasts to represent the subseasonal time scale. Biases, however, are qualitatively similar for the other forecast weeks. For both subregions and all models, forecast probabilities tend to be generally very low, consistent with the extreme nature of TCs, leading to low relative frequencies of TC occurrence in the target variable. Thus, the model predictions can be made only with low confidence as the forecast probabilities are distributed mainly around the mean relative frequency of the target variable. However, it can be stated that the logistic regression models can predict with slightly higher confidence compared to the benchmark models.

Recall that a model is well-calibrated (or reliable) when the forecast probabilities match the observed relative frequencies. Miscalibration can thus be visually assessed through deviations

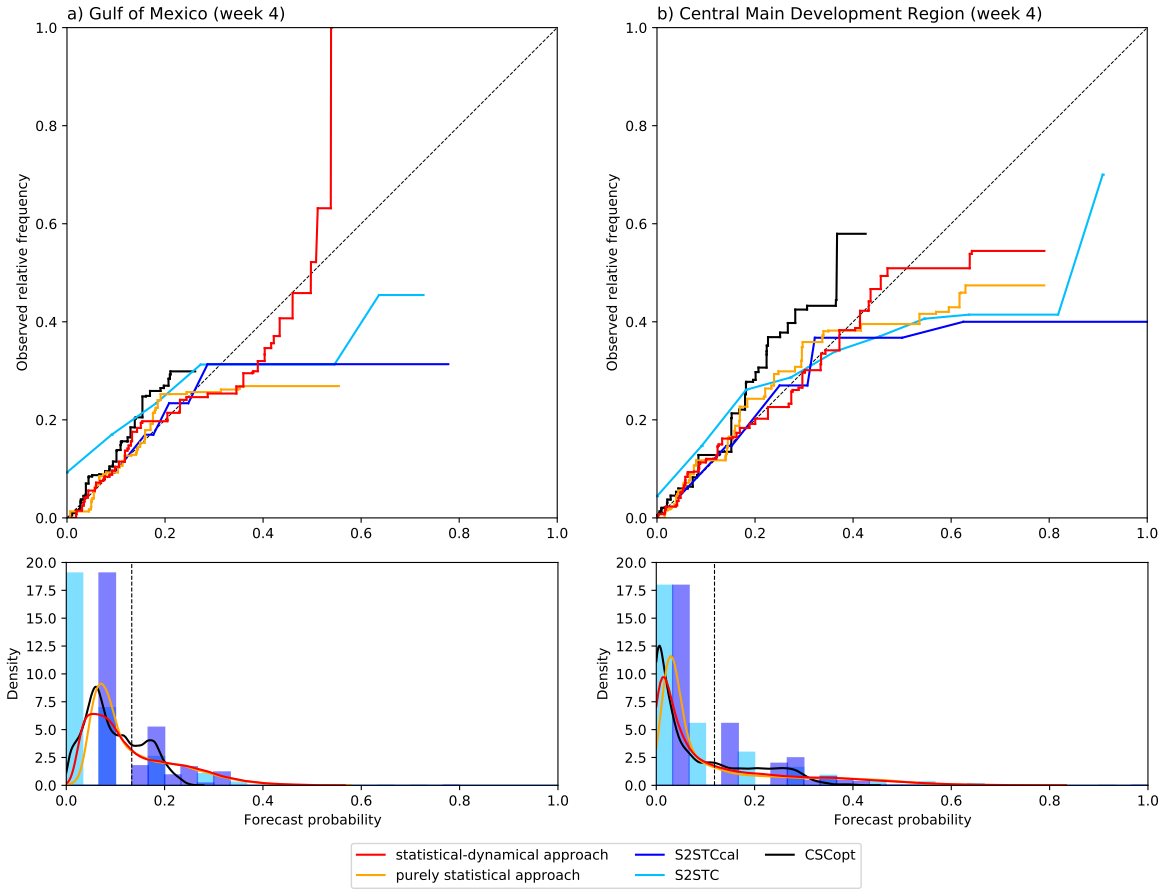

Figure 6.6: CORP reliability diagram for Gulf of Mexico (a) and central MDR (b) week-four forecasts, respectively. While probability forecast distributions are visualized by means of histograms for the S2STC and S2STCcal models, a kernel density estimation is applied to generate continuous curves for the other models. The dashed vertical line indicates the mean relative frequency of the target variable.

of the calibration curves from the diagonal. The first thing to notice is that all models are more reliable for low forecast probabilities than for higher ones, which is consistent with the refinement distributions discussed before. The underforecasting situation (TC nonoccurrence bias) of the CSCopt model is likely to result from a reduced TC occurrence in the 1968–1997 period, which was used to extend the 1998–2017 validation period for calculating more robust climatologies. However, since the CSCopt is also a base predictor for the logistic regression models, it has no competitive disadvantage when evaluating model skill. The S2STC model similarly underforecasts the low forecast probabilities, but overforecasts the few high forecast probabilities, which results in a general overconfidence. To correct for this conditional bias, this particular NWP-based model is calibrated using IDR, as described above. The S2STCcal follows the diagonal quite well for low forecast probabilities, and thus generates much more reliable forecasts. Since logistic regression is known to yield well-calibrated

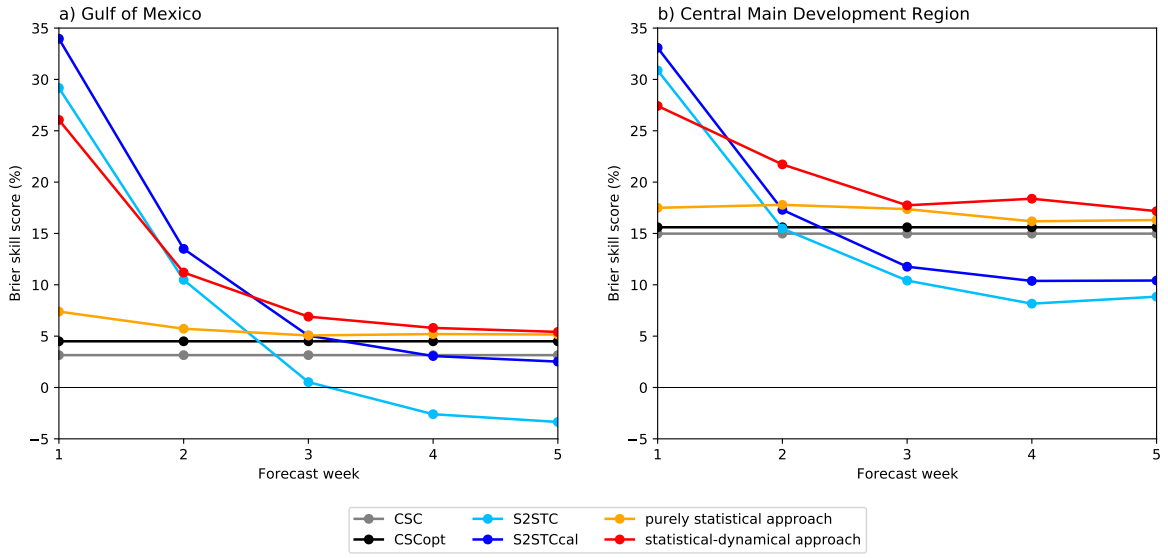

Figure 6.7: Brier skill score (BSS; %) as a function of forecast week for the CSC (gray), CSCopt (black), S2STC (lightblue), S2STCcal (darkblue), purely statistical (orange), and statistical-dynamical (red) models, respectively, relative to the MSC and validated in the Gulf of Mexico (left) and central MDR (right) subregions.

forecasts, the calibration curves for the two approaches of logistic regression models are well-aligned with the diagonal for low forecast probabilities. The increasing deviations with higher forecast probabilities are likely due to the few samples, which are obviously insufficient for generalization. Overall, subseasonal forecasts of the logistic regression models, with a slightly better calibrated statistical-dynamical approach for higher forecast probabilities, are more reliable than the benchmark forecasts.

Figure 6.7 shows a comparison of the BSS as a function of forecast week for the different model types validated in the Gulf of Mexico and central MDR subregions. Since climatological forecasts are independent of the forecast week, the BSS also does not change with lead time. Considering that the MSC is used as reference, the positive BSS for the CSC and CSCopt models indicate that the ability to simulate seasonal variations is rewarded. The improvement in skill, however, exhibits remarkable subregional differences, as can be seen by a BSS of CSC three times stronger (about 15% vs. less than 5%) for the central MDR compared to the Gulf of Mexico. The BSS of the CSC is further enhanced when correcting for the undersampling problem of the CSC through a locally optimized smoothing. The relative enhancement is found to be much stronger for the Gulf of Mexico than the central MDR subregion, which can be explained by the more variable CSC.

In terms of the NWP-based benchmark models, IDR calibration helps increase the BSS of S2STC by adding 3–6% and 1–2% for the Gulf of Mexico and the central MDR, respectively, over the forecast weeks considered. For forecast week one, the S2STCcal model by far exceeds the CSCopt, but rapidly looses most of its skill over the first two forecast weeks, i.e., on

the medium range, eventually leveling off thereafter on subseasonal timescales. While the CSCopt outperforms the S2STCcal from week three on in the central MDR, the CSCopt takes the lead only beyond forecast week three in the Gulf of Mexico. Apart from these minor subregional differences, this considerable drop in model skill around week two to three is in accordance with previous findings for forecasts of basin-wide TC occurrence (Lee et al., 2018), highlighting the potential of climatological forecasts for subseasonal timescales.

Expanding the climatological model by including predictors generated from past data, the purely statistical approach improves the CSCopt skill for all five forecast weeks. While 3% are added in the Gulf of Mexico at week one, improvements reduce to less than 0.7% beyond medium range (Fig. 6.7a). In comparison, a maximum of 2% is added to the CSCopt BSS in the central MDR, but this level of improved skill drops to about 0.7% only after week three (Fig. 6.7b).

Replacing the past data with the S2S ensemble mean and standard deviations for each predictor, the statistical-dynamical approach further raises the BSS at all forecast weeks. The gain in skill is greatest for week one, and continuously decreases with longer lead times, except for minor subseasonal variations in the central MDR. For the Gulf of Mexico, improvement in skill from the purely statistical to the statistical-dynamical approach is 4.5–6.5 (0.4–3.2) times greater on the medium (subseasonal) range than the improvement from the CSCopt to the purely statistical approach. In analogy, for the central MDR, relative improvements appear to be 1.8–5.2 (0.2–3.8) times larger on the medium (subseasonal) range. Even though both logistic regression models are beaten by the S2STCcal model at week one, they outperform all benchmark models from week three (two) on in the Gulf of Mexico (central MDR). Note that a simple approach to obtain at least equivalent skill for week one and two would be to include the S2STCcal forecasts as a predictor to the logistic regression models.

Summary and conclusions

Keeping in mind the aim of this section, that is, to give a motivation for the use of hybrid forecast models, we summarize the findings in the following:

- While the S2S ECMWF model predicted best at week one, it quickly dropped in skill thereafter due to the chaotic nature of the atmosphere blurring the valuable information contained in the initial conditions. Analogous to the results in Chapter 5, postprocessing of the underforecasting S2STC model via IDR helped to raise skill at all lead times, but did not exceed the other approaches on subseasonal timescales.

- The purely statistical approach from Leroy and Wheeler (2008) and Slade and Maloney (2013), with logistic regression models trained on past data predictors, improved skill over the CSCopt model in both subregions out to week five.

- With the statistical-dynamical approach, an even greater increase in model skill was found at all lead times considered, but especially on the medium range. Though this

approach was still worse than the S2STCcal for week one, despite a significant increase in skill over the purely statistical approach, it outperformed all other models in the Gulf of Mexico (central MDR) from week three (two) on. In the Gulf of Mexico, the subseasonal improvement from the purely statistical to the statistical-dynamical approach is 0.4–3.2 times larger as the one from the CSCopt to the purely statistical approach. The analogous for the central MDR yields a factor for relative improvement of 0.2–3.8. In view of the generally lower CSC skill in the Gulf of Mexico, such an improvement becomes even more remarkable, highlighting the value of this approach for subregions that are less subject to a seasonal cycle.

The systematic comparison of original and hybrid model types presented in this section has demonstrated the great potential of statistical-dynamical modeling for a specific application of extreme events on the subseasonal forecast horizon. Exploiting S2S forecasts to develop such hybrid models proved to be the best strategy, at present, for probabilistic forecasting of subregional North Atlantic TC occurrence beyond week two, and might be a promising strategy for other forecasting applications as well.

The conclusions we draw from this section are that hybrid models can be used to leverage predictive performance. Here, domain knowledge is used to combine dynamical NWP forecasts with statistical models based on climatological data. In the following, we want to combine our NN-based postprocessing with an identification of the high-wind areas, resulting in a hybrid model that is hopefully also able to leverage predictive performance by incorporating domain knowledge in a standard forecasting model.

6.3 Identification of high-wind features within winter storms

In the mid-latitudes, extratropical cyclones can produce some of the most severe natural hazards, especially during wintertime. These winter storms can cause high wind speeds, heavy precipitation, storm surges and, thus, considerable damage. High winds are typically associated with four mesoscale features within the synoptic-scale cyclone, namely the *warm (conveyor belt) jet* (WJ), the *cold (conveyor belt) jet* (CJ), *cold-frontal convective gusts* (CFC), *strong cold sector winds* (CS) and, at least in some storms, the *sting jet* (SJ). All features can cause damage due to strong gusts, such that it is important to accurately forecast them and their associated wind fields. In Section 6.1, we saw that even sophisticated postprocessing methods are subject to forecast busts in these winter storms, hence, fail to achieve this objective. By developing an objective identification algorithm for the wind features, this section lies the foundation for further exploring the idea of a feature-dependent postprocessing method, which was proposed in Section 6.1.

As previously proposed approaches for the identification of the wind features are purely

subjective and relatively time consuming (Parton et al., 2010; Hewson and Neu, 2015; Earl et al., 2017), and thus hard to automate, we aim to develop an objective analysis of the different mesoscale wind features that can flexibly be applied to station and gridded data and, therefore, serve as a basis for climatological studies, forecast evaluation and postprocessing development. The strategy we follow is to start with a subjective identification (as in previous studies) but to use the results to then train a probabilistic RF to develop an objective procedure that can be applied to cases outside of the training dataset. The identification is designed to be independent of horizontal gradients, hence resolution, and can principally be applied to observations from a single weather station. In addition, the identification is based on tendencies over 1 hour only, making it applicable to time series with gaps. Our new developed method is referred to as RAMEFI (*RAndom forest-based MEsoscale wind Feature Identification*). Given that the provision of a feature-dependent postprocessing tool can enhance the forecasts of strong winds and wind gusts by eliminating the forecast busts observed in Section 6.1, it can potentially contribute towards better weather warnings and impact forecasting of such events (e.g., Merz et al., 2020). In this section, we will show examples using surface stations and COSMO reanalysis data. The output of the RF are feature probabilities rather than a binary identification, which allows an evaluation of how well individual data points fit the typical feature characteristics and the identification of joint features or transition zones.

The section is structured as follows. First, the used datasets and methods are discussed in Section 6.3.1. Section 6.3.2 then details our new method RAMEFI, starting from the subjective labeling of wind features through the training of the RF to the display of areal feature information. The performance of the RF probabilities is then assessed in Section 6.3.3, followed by the conclusions.[11]

6.3.1 Data and method

Here, we introduce the observational and model data as well as 12 winter storm case studies used for the training and evaluation of RAMEFI. Furthermore, it describes how we assess probability predictions obtained by the RF.

Surface observations

The main basis of our analysis is a dataset of hourly surface observations from 2001 to mid 2020. This includes mean sea level pressure (p), air temperature at 2 m (T), wind speed at 10 m (v), wind direction at 10 m (d) and precipitation amount (RR). Using T and p, we further compute the surface pressure using the barometric height formula to then calculate the potential temperature (θ). Our focus is on Europe, more specifically, stations within

[11] Note that the publication underlying this section is a collaboration with meteorologist (see Section 1.1 for my contributions).

Table 6.2: Overview of the variables considered for the objective identification using the probabilistic RF. The fourth column indicates whether the variable is used as predictor variable for the final version of the RF. The associated percentiles and medians are computed with respect to the location, time of day and day of the year ±10 days.

Variable	Description	Unit	RF	Derivation
v	Wind speed at 10 m	m/s	-	Station observation
$\tilde{v}$	Normalized wind speed	Unitless	✓	$\tilde{v} = v/v_{98}$, where v_{98} is the associated 98th percentile
d	Wind direction	°	✓	Station observation
Δd	Tendency of wind direction	°/h	✓	$\Delta d = d - d_{-1}$, where d_{-1} is the observation of the previous hour
p	Mean sea level pressure	hPa	✓	Station observation
Δp	Tendency of mean sea level pressure	hPa/h	✓	$\Delta p = p - p_{-1}$, where p_{-1} is the observation of the previous hour
T	Air temperature at 2 m	K	-	Station observation
θ	Potential temperature	K	-	Derived from T and p
$\tilde{\theta}$	Normalized potential temperature	Unitless	✓	$\tilde{\theta} = \theta/\theta_{50}$, where θ_{50} is the associated median
$\Delta\tilde{\theta}$	Tendency of normalized potential temperature	1/h	✓	$\Delta\tilde{\theta} = \tilde{\theta} - \tilde{\theta}_{-1}$, where $\tilde{\theta}_{-1}$ is the derived value of the previous hour
RR	Precipitation	mm/h	✓	Station observation

the area of −10 to 20° E, 40 to 60° N. Around 1,700 stations are included; however, less than 400 of these stations observe on average all five parameters. For the training of the RF (Section 6.3.2), we focus on stations that measure at least three of the five parameters. The most frequent missing parameter in the hourly data is RR, as many stations only measure 3- or 6-hourly precipitation. However, many stations, especially over Germany, measure RR only and, hence, are not usable for the training of the RFs but still helpful to inform our subjective labeling. In addition, we exclude mountain stations, i.e., those with a station height above 800 m, as we suspect these to be dominated by orographic influences that may blur the feature characteristics we want to identify. This leaves around 750 stations per time step.

To take into account the diurnal and seasonal cycles as well as location-specific characteristics (e.g., exposed stations in coastal regions) in θ and also v, we decided to normalize these parameters by their climatology. For θ, this means $\tilde{\theta} = \theta/\theta_{50}$, where θ_{50} is the median for the specific location, time of day and day of the year ±10 days. This is done analogously for v using the 98th percentile v_{98} and $\tilde{v} = v/v_{98}$, as we are mostly interested in high winds in this section. The 98th percentile is used in analogy to standard high-wind quantities such as the SSI, which is computed from stations where measured gusts exceed the local 98th percentile and provides an integral indication for the strength of the cyclone and the associated potential damage. Both θ_{50} and v_{98} are computed for the time period 2001 to 2019. Moreover, we are interested in temporal tendencies of p, $\tilde{\theta}$ and d, here represented simply by the difference between the current and the prior time step (Δp, $\Delta\tilde{\theta}$ and Δd, respectively). All parameters and their descriptions are listed in Table 6.2.

COSMO-REA6

As an example for a gridded dataset, we use COSMO-REA6 data from the Hans-Ertel-Centre for Weather Research, which is a reanalysis based on the COSMO model from DWD covering the European CORDEX domain with a grid spacing of 0.055°, i.e., roughly 6 km (Bollmeyer

Table 6.3: Selected winter storm cases from 2015 to 2020 over central Europe (names as given by Free University of Berlin), date, maximum observed gust speed (location), SSI over Germany and associated high-wind features.

Case	Date	Maximum observed gust speed		SSI	Features
		above 800 m	below 800 m		
Niklas	31 March 2015	192 km/h (Zugspitze, D, 2964 m)	148 km/h (Weinbiet, D, 553 m)	20.8	WJ, CFC, CJ
Susanna	09 February 2016	158 km/h (Patscherkofel, AT, 2247 m; Pilatus, CH, 2106 m)	158 km/h (Île de Groix, FR, 46 m)	3.6	WJ, CFC, CJ
Egon	12–13 January 2017	150 km/h (Fichtelberg, D, 1231 m)	148 km/h (Weinbiet, D, 553 m)	5.9	WJ, SJ, CJ
Thomas	23–24 February 2017	158 km/h (Brocken, D, 1134 m)	152 km/h (Capel Curig, UK, 216 m)	3.0	WJ, CJ
Xavier	05 October 2017	202 km/h (Sněžka, CZ, 1602 m)	141 km/h (Saint-Hubert, BE , 563 m)	6.3	WJ, CJ
Herwart	29 October 2017	176 km/h (Fichtelberg, D, 1231 m)	144 km/h (List/Sylt, D, 26 m)	15.2	WJ, CFC
Burglind	03 January 2018	217 km/h (Feldberg, D, 1490 m)	150 km/h (Wädenswil, CH, 463 m)	15.2	WJ, CFC, CJ
Friederike	18 January 2018	204 km/h (Brocken, D, 1134 m)	144 km/h (Hoek Van Holland, NL, 7 m)	18.3	WJ, SJ, CJ
Fabienne	23 September 2018	141 km/h (Feldberg, D, 1490 m)	158 km/h (Weinbiet, D, 553 m)	4.6	WJ, CFC
Bennet	04 March 2019	181 km/h (Cairngorn, UK, 1245 m)	151 km/h (Cape Corse, FR, 106 m)	5.1	WJ, CFC
Eberhard	10 March 2019	194 km/h (Sněžka, CZ, 1602 m)	141 km/h (Weinbiet, D, 553 m)	10.1	WJ, CJ
Sabine	09–10 February 2020	195 km/h (Sněžka, CZ, 1602 m)	219 km/h (Cape Corse, FR, 106 m)	20.0	WJ, CFC

et al., 2015).[12] The reanalysis is available from 1995 to 2019. This means that one of our case studies, namely storm Sabine, is not included (Table 6.3). The same surface parameters as for the observational data are used. The dataset contains p, T, RR and the zonal and meridional wind components, from which we can compute v and d. Again, we further calculate $\tilde{\theta}$ and the temporal tendencies Δp, $\Delta\tilde{\theta}$ and Δd. Due to computational cost, we compute θ_{50} and v_{98} for the 10-year time period from 2005 to 2015 only, but this should have a negligible effect on the final outcome.

Case studies

In this section, we focus on 12 winter storm case studies between the years 2015 and 2020 listed in Table 6.3. Analogous to the winter storms selected for postprocessing in Section 6.1, the selection was based on the SSI over Germany, caused damage and impacted area.[13] This includes the eight winter storms with the highest SSI during this time period plus four subjectively chosen more moderate storms to capture a healthy diversity of cyclones and features. The selected cases occurred during the extended winter half-year between the end of September and end of March. They vary in terms of their cyclone tracks and occurring high-wind features. Two case studies developed SJs, namely Egon (Eisenstein et al., 2020) and Friederike. We also include two storms, named Herwart and Sabine, with an exceptional large pressure gradient leading to a stronger background wind field, such that it is more

[12] CORDEX refers to the COordinated Regional climate Downscaling EXperiment EUR-11 domain (Giorgi et al., 2009).

[13] Note that the SSI values of Niklas and Susanna differ from those listed in Table 6.1. The SSI values deviate because the underlying datasets (and thus surface stations) also deviate. Due to this property, the SSI is in general compared only when calculated consistently, e.g., within a study.

difficult to distinguish the features and the contribution of them to the storm's wind footprint. Further, Sabine stands out to be an extremely deep cyclone with a minimum core pressure of 944 hPa during its lifetime.[14] This is atypical for winter storms, and should be considered in the statistical evaluation.

Assessing probability predictions for multiple wind features

Probability predictions of three or more classes, such as the wind features, are typically evaluated by downscaling to two-class problems, of which the *one-against-all* and *all-pairs* approaches are two well-known examples (Zadrozny and Elkan, 2002). While the one-against-all approach compares the occurrence of one wind feature against all others grouped together, the all-pairs approach considers the conditional probabilities for each pair of classes, for example, the conditional probabilities of the WJ and the CJ when one of the two features materializes. The one-against-all approach is used to evaluate how well one specific wind feature is forecast, the all-pairs approach to evaluate the ability to discriminate between two wind features.

The probabilities are evaluated based on the paradigm that a prediction should aim to maximize sharpness subject to calibration. As defined in equation (2.2), a probability forecast f is called calibrated if the *conditional event probability* (CEP) matches f. Further, a probability prediction is said to be sharper, the more confident the prediction is, that is, the closer to 0 or 1. We will assess the calibration of the probability forecasts qualitatively via CORP reliability diagrams. In addition to the calibration curve, the frequency of the probabilities is illustrated by a histogram. The more U-shaped the histogram is, the closer the predictions are to 0 and 1 and thus the sharper. Quantitatively, calibration and sharpness are assessed using the BS and the BSS with respect to the class frequencies observed in the training data of the RF.

Consider a multiclass probability prediction $\mathbf{p} = (p_1, \dots, p_K)$, where $p_1, \dots, p_K \in [0, 1]$ and $\sum_{i=1}^{K} p_i = 1$, for a nominal target variable $Y \in \{1, \dots, K\}$, with $K \geq 3$ classes that are not ordered. Analogous to the calibration criterion for binary probability predictions in equation (2.2), $\mathbf{p}$ is called *(auto-)calibrated* if $\mathbb{Q}(Y = i \mid \mathbf{p}) = p_i$ almost surely for all $i = 1, \dots, K$ (Gneiting and Ranjan, 2013). The multivariate version of the BS for a probability vector $\mathbf{p}$ and realizing class $i \in \{1, \dots, K\}$ is given by $S(\mathbf{p}, i) = \sum_{j=1}^{K} (p_j - \mathbb{1}\{i = j\})^2$ (Brier, 1950).

The one-against-all approach reduces the multiclass prediction problem to a set of K dichotomous problems. For each class $i \in \{1, \dots, K\}$, the probability p_i is a prediction for $\tilde{Y} = \mathbb{1}\{Y = i\}$. Note that evaluating the predictions p_i for $\tilde{Y}$ for each class is not equivalent to checking the multiclass calibration criterion, as the joint distribution of p_i and not that of $\mathbf{p}$ is considered in the one-against-all approach. The all-pairs approach reduces the multiclass prediction problem to a set of $K(K-1)/2$ dichotomous problems. For each pair of classes

[14] The term deep refers to the associated mean sea level pressure.

(i, j), where $i, j \in \{1, \ldots, K\}$ and $i > j$, we consider only samples with $Y \in \{i, j\}$. Then, the conditional probability $\tilde{p}_{i,j}$ is a prediction for $\tilde{Y}$, where

$$\tilde{p}_{i,j} = \frac{p_i}{p_i + p_j} \quad \text{and} \quad \tilde{Y} = \begin{cases} 1 & \text{for} \quad Y = i, \\ 0 & \text{for} \quad Y = j. \end{cases} \tag{6.4}$$

6.3.2 RAMEFI method

Our new method, RAMEFI, focuses on strong but not exceptionally high wind speeds. These wind speeds are usually indicated by the 98th percentile. To obtain a sufficiently large storm area and to base that on a widely used reference, we decided to include stations reaching 80% of their 98th percentile, i.e., $\tilde{v} \geq 0.8$. To capture usually narrow and fast-moving features such as CFC, RAMEFI requires hourly data. All used parameters are independent of the location of the station/grid point and horizontal gradients, such that, in principle, the approach can be applied to a single station and datasets with differing horizontal resolution. The approach evaluates each 1 hour interval independently.

RAMEFI includes three steps described in the following subsections. First, we identify the features subjectively in surface observations in 12 selected case studies, such that each station is assigned to a specific feature. These labels are then used to train RFs for feature prediction on the basis of a CV approach. In a final step, we obtain forecasts on a grid by interpolating the predicted probabilities using a Kriging approach. For the COSMO-REA6 data, the features are identified analogously. Instead of training separate RFs, we apply the RFs trained on the surface observations. As the COSMO-REA6 forecasts are already grid-based, the Kriging step is obsolete.

(1) Subjective labeling

Given the sometimes unclear distinction between the high-wind features of interest in realistic cases, we decided to base our algorithmic development on how experienced meteorologists would identify the features on the basis of a wide range of parameters and their evolution in time and space. The guiding principles for the labeling were extracted from the scientific literature and are mainly based on the location relative to the cold front and cyclone core. In our surface parameters, a cold front is then mostly identified through the characteristic change of the sign of Δp. It is labeled CFC, if a larger area of precipitation along it is observed, while high winds ahead of the front within the warm sector are labeled WJ. The CJ is mostly detected through its hook-shaped wind footprint at the tip of a wrapped-around occlusion or bent-back front as well as through its proximity to the cyclone center. An SJ is labeled when model-based trajectories analogous to Eisenstein et al. (2020) confirm a descending airstream. The area behind the cold front that is not associated with the CJ or SJ is labeled as the CS.

The subjective labeling was done for the introduced 12 case studies (Section 6.3.1). In total,

282 time steps have been analyzed. As mentioned above, we excluded mountain stations and stations where less than three of the given parameters were measured. This leaves around 750 stations per time step for the subjective labeling. Overall, for the 12 case studies, we have 77,517 data points where $\tilde{v} \geq 0.8$, of which 1,200 (24.77%) are not associated with a feature (NF), 21,809 (28.13%) were labeled as CS, 19,501 (25.16%) as WJ, 11,705 (15.1%) as CJ, 3,800 (4.9%) as CFC and 1,502 (1.94%) as SJ. However, the SJ is a small, short-lived and rare feature, and the characteristics of SJs and CJs in surface parameters are very similar due to the proximity in both time and space. A first training with SJ and CJ as separate features showed that a clear distinction is not possible with the information at hand and that the SJ is mostly detected as CJ. Therefore, we decided to include it in the more frequent CJ feature, increasing the values for CJ to 13,207 data points (17.04%).

The features were further labeled in all case studies (except for Sabine, which occurred outside of the reanalyis time period) for COSMO-REA6 data. These labels are used to evaluate the predictions generated by the station-based RFs for a grid-based dataset (Section 6.3.3). For computational reasons, i.e., as labels are set for every grid point rather than an area, we downsampled the COSMO grid to every third grid point in the zonal and meridional directions, resulting in a grid spacing of 0.1875° (around 21 km). Moreover, we excluded ocean grid points, as the characteristics of the high-wind features might be different from land due to different surface friction and surface heat fluxes, among other factors. Regions with a high wind speed not directly associated with a winter storm, especially over Italy and the Balkans, were not labeled.

(2) Probabilistic Random Forest

RF (Breiman, 2001) is a popular, robust ML method for classification and regression problems that does not rely on parametric assumptions but instead is based on the idea of decision trees (Breiman, 1984). In Section 4.2.5, where we introduced the QRF technique for statistical postprocessing, we have already lined out the RF approach. While QRF generates (a set of) quantile predictions for the target variable, we here obtain probability forecasts by using the frequencies of the observed wind features among the samples in the corresponding leaf. In a meteorological context, probabilistic RFs have already been applied to predict damaging convective winds (Lagerquist et al., 2017) and severe weather (Hill et al., 2020).

ML methods such as the RF are often referred to as black boxes due to a lack of interpretability, although there exist several techniques to understand what the models have learned and how the predictions are related to the predictor variables, as demonstrated in Section 5.3.3, where we found that the statistical postprocessing methods based on ML learn physically consistent relations. We will apply two predictor importance techniques, one to find the most relevant predictors and one that illustrates the effect of the predictor values on the RF probabilities. The first is the permutation importance of a predictor, which has been described in Section 5.3.3 of the case study on wind gust prediction. Recall that, proceeding

separately for each predictor, the values of that predictor are shuffled randomly within the test data in space and time such that the physical relation to the observed wind feature is broken. Then, based on these permuted predictor values, new predictions are generated and compared to those obtained with the original data. The worse the predictions become (with respect to an evaluation measure), the more important the predictor. Here, we measure predictive performance with the BS, the importance measure is referred to as BS permutation importance. The second technique is the *partial dependence plot* (PDP; Greenwell, 2017), which illustrates the effect of a predictor on the prediction. Given a fixed predictor, a PDP shows the expected RF probability dependent on the value of the predictor variable while averaging out the effects of the other predictors. Hence, a PDP illustrates how the RF probabilities depend on the value of a specific predictor variable, on average. For more details, we refer to McGovern et al. (2019).

For RAMEFI, we apply RFs to generate probabilities of the high-wind features. The predictor variables used are listed in Table 6.2. For the station-based observations, we use a CV scheme on the different winter storm cases, that is, for each winter storm, the predictions are generated by an RF that is trained on the data of the remaining 11 winter storms. Training RFs in a similar CV scheme for the COSMO-REA6 data becomes computationally infeasible, as the underlying datasets become too large. Since the underlying processes should coincide for both the station- and model-based data, we instead apply the station-based RFs in the same scheme to generate probabilities using the COSMO-REA6 data. For the PDPs, one partial dependence curve has to be calculated for each RF generated in a fold of the CV, that is, for each winter storm. The final curves are then obtained by a weighted average depending on the sample size of the folds.

Details on the implementation, including the choice of the hyperparameters, are described in the following. Analogous to QRF in Section 5.3, RF is implemented via the `ranger` package (Wright and Ziegler, 2017). Table 6.2 summarizes the predictors used, Table 6.4 the chosen hyperparameters. One question in the implementation is the handling of missing values, which an RF cannot process. The station-based samples frequently miss values of one or more predictor variables, especially precipitation is affected. We tried different strategies to handle missing values such as leaving out instances with missing values or replacing the missing values with a mean value, and found similar results. Therefore, we decided to replace the missing values in order to use the largest sample size possible, which is desirable for the evaluation and the Kriging step. In each fold of the employed CV scheme, the missing values (both in the training and test set) are replaced by the mean value of the associated predictor variable in the training set.

Due to normalizing θ and v, the trained RF is fairly independent of location-specific information, such that it can hopefully be applied successfully to other midlatitude regions around the world affected by extratropical cyclones. However, before doing that, we recommend a thorough sanity check, particularly when using it over the ocean and mountainous regions.

Table 6.4: Overview of the hyperparameters of the probabilistic RF.

Hyperparameter	Value
Number of trees	1,000
Number of predictors considered at each split	2
Minimal node size	10
Maximal depth	Unlimited
Splitting criterion	Gini

(3) KRIGING

As it is difficult to envision a coherent area of a certain wind feature from probabilities at single stations that are distributed irregularly over the study area, we interpolate the station-based probabilities to a regularly spaced grid in order to visualize the results. In geostatistics, this is generally achieved by *Kriging* (Matheron, 1963). In principle, the Kriging predictions (here on the grid) are the weighted averages of the input data (here the station data), where the specification of the weights is driven by the covariance of the underlying random process. Under the assumption of Gaussianity (Rasmussen and Williams, 2005), Kriging provides the optimal full predictive distribution. The key requirement for the implementation of Kriging in the context of Gaussian processes is the specification of the mean and the covariance function.

Here, we perform univariate Kriging to obtain probability maps for each wind feature, where we specify the mean and covariance function by a constant mean function and the stationary Matérn covariance function (Matérn, 1986; Guttorp and Gneiting, 2006). For the estimation, we resort to MLE for Gaussian processes. However, as the input data is, in our case, probabilities and thus deviates from the Gaussianity assumption, we perform a data transformation for approximate Gaussianity. For the production of probability maps, we independently perform Kriging on each of the class probabilities (hence univariate Kriging) and normalize the resulting probabilities for each grid point such that, across the multiple wind feature, the probabilities sum to one. Note that the Kriging predictions are only obtained for areas over land, where our winter storms occurred and where a sufficient amount of data was available for a reliable interpolation.

Next, we provide a brief mathematical formulation of the Kriging approach. Let $\{X(\mathbf{s}),\ \mathbf{s} \in \mathbb{R}^2\}$ be the spatial Gaussian process modeling the transformed probability of a certain wind feature, indexed by the spatial coordinates $\mathbf{s}$ that correspond to the latitude and longitude associated with the (transformed) probability. Further, we denote the mean function by $\mathbb{E}\{X(\mathbf{s})\} = \mu(\mathbf{s})$ and the covariance function by $\text{Cov}\{X(\mathbf{s}), X(\mathbf{s}')\} = C(\mathbf{s}, \mathbf{s}')$. Then, for a given set of station-based data $\boldsymbol{x} = \{X(\mathbf{s}_1), \ldots, X(\mathbf{s}_n)\}^T$, the spatial prediction at a grid cell $\mathbf{s}_0$ is given as $\hat{X}(\mathbf{s}_0) = \mu(\mathbf{s}_0) + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$, where $\boldsymbol{\mu} = \{\mu(\mathbf{s}_1), \ldots, \mu(\mathbf{s}_n)\}^T$, $\Sigma_{12} = \{C(\mathbf{s}_0, \mathbf{s}_1), \ldots, C(\mathbf{s}_0, \mathbf{s}_n)\}$ and $\Sigma_{22} = \{C(\mathbf{s}_i, \mathbf{s}_j)\}_{i,j=1}^n$. Additionally, one can obtain the prediction variance as $\text{Var}\{\hat{X}(\mathbf{s}_0)\} = C(\mathbf{s}_0, \mathbf{s}_0) - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T$, and the full predictive

distribution as $X(\mathbf{s}_0) \mid \boldsymbol{x} \sim \mathcal{N}(\hat{X}(\mathbf{s}_0), \sqrt{\text{Var}\{\hat{X}(\mathbf{s}_0)\}})$. The choice and estimation of the mean function μ and the covariance function C are key elements of the Kriging implementation. While one can choose any parametric or nonparametric functional representation for μ, the valid choice for C is limited to the class of positive semidefinite functions. In practice, the covariance function is often assumed to be stationary, which implies that the covariance function depends on the spatial locations only through spatial lags, i.e., $C(\mathbf{s}_i, \mathbf{s}_j) = K(\mathbf{s}_i - \mathbf{s}_j)$ for some positive semidefinite function K. In our implementation, we have specified the mean function $\mu(\cdot) = c$, $c \in \mathbb{R}$, to be a constant valued function, and the covariance function K to be the Matérn class of stationary covariance function (Matérn, 1986).

In our practical implementation of Kriging, we transform the probabilities by using the `bestNormalize` package (Peterson, 2021) to achieve approximate Gaussianity, which automatically chooses a suitable transformation from a set of commonly used transformations. The probabilities on the grid generated via the univariate Kriging need to be normalized such that they sum up to 1. However, at some grid cells distant from the cyclone track, the predicted probabilities are small for all of the wind feature and normalization results in unrealistic predictions. Thus, we only perform the normalization at grid cells where the accumulated probability is larger or equal to 20%. For the visualization, we further drop the grid cells where the largest normalized probability is smaller than 20% (which includes the grid cells for which no normalization was performed).

6.3.3 Results

In Section 6.3.1, we described how we evaluate probability predictions for the wind features. Here, we first apply this concept to the RF probabilities for the station data and the COSMO reanalysis. Then, we investigate the relationship between the predictors and the RF probabilities. The section ends with a discussion of the advantages and shortcomings of not using spatial dependencies in the feature identification.

Evaluation of the RF probabilities

The evaluation of the station-based RF probabilities is split into three parts. First, we quantitatively compare the RF forecasts with the class frequencies in the training data, then we assess how well the RFs predict the individual wind features in the one-against-all approach and lastly we check how well the predictions distinguish two features with the all-pairs approach. For each storm that we predict, the class frequencies of the other 11 storms are used as a benchmark prediction. As expected, we find that the RF probabilities outperform the benchmark in terms of the (multivariate) BS for the prediction of each winter storm. The overall improvement is 24.7%, while for the different storms it ranges from 11.8% to 34.7% with 11.8% being the skill for Xavier, which is discussed in some detail at the end of the section.

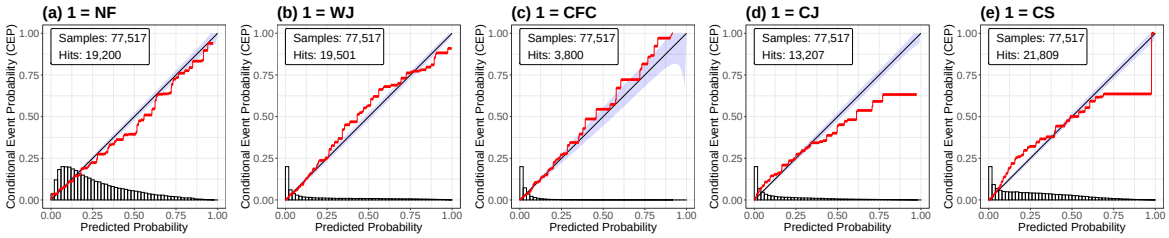

Figure 6.8: CORP reliability diagrams of the RF probabilities for the individual wind features in the one-against-all approach including all 12 storms.

Figure 6.8 shows the reliability diagrams of the RF probabilities in the one-against-all approach for the occurrence of NF and the four specific wind features (WJ, CFC, CJ and CS). We observe that the probabilities are in general well-calibrated for all five cases, as the calibration curves closely follow the diagonal. The predictions are generally reliable, especially for small probabilities, which are most frequent in this setting, as the peaks of the histograms illustrate. Therefore, the RFs identify the nonoccurrence of a specific wind feature with high confidence (Figure 6.8a). For larger probabilities, the predictions of NF, the WJ and the CFC are well-calibrated, as the calibration curves stay reasonably close to the diagonal (Figure 6.8a–c), while for the CJ and CS (Figure 6.8d,e) larger deviations are evident. In both cases, the RF overforecasts the events, that is, the predicted probability is generally too large.

The reliability diagrams of the all-pairs approach are displayed in Figure 6.9, which show that the RFs yield well-calibrated probabilities for the distinction of all feature pairs but one. When the RF predicts that the CJ is more likely to occur than the CS (in case one of those two materializes), the RFs overforecast the CJ, meaning that the CS occurs more often than predicted (Figure 6.9j). This is consistent with the results from the one-against-all approach, where we found that the CJ and CS predictions were not well-calibrated for high probabilities, indicating that the RF fails to distinguish them for large conditional probabilities of the CJ. Further, the histogram of this pairwise comparison shows that the RF cannot discriminate between the two features with high confidence. This issue can be seen best for the storms Herwart and Sabine, which both did not develop a CJ, although a CJ was identified by the RFs. The main meteorological reason for this problem is the general similarity of the two features and that the hook-shaped structure, which is used for the subjective identification of a CJ, cannot be considered in the RF, such that the distinction is mainly based on p, as will be discussed in the predictor importance part of this section. Other than that, the calibration curves of the other pairs follow closely the diagonal. Moreover, we note that the WJ is distinguished well from the CJ and CS, as the U-shaped histograms of the probability distributions show (Figure 6.9f,g).

For the predictions derived from the COSMO-REA6 data, the RF probabilities are also able to distinguish the features well, although the RFs used were trained on station-based data. The predictions exhibit similar characteristics and perform (as expected) only slightly

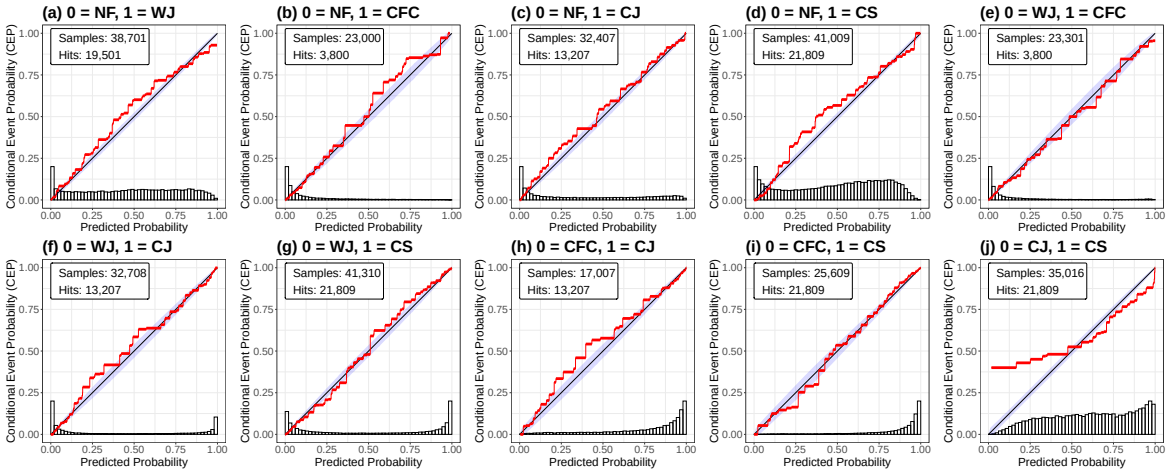

Figure 6.9: CORP reliability diagrams of the conditional RF probabilities comparing two wind features in the all-pairs approach including all 12 storms.

worse than for the station data. As before, the skill of the BS is calculated with respect to a benchmark prediction based on the class frequencies and is 19.6% for all storms. For eight of the selected storms, we observe improvements ranging from 11.0% to 37.5%; however, for Herwart and Susanna, the skill scores are -0.8% and -11.6%, respectively, indicating a decrease in predictive performance. For Susanna, this is due to a larger high-wind region ahead of but not directly connected to the cyclone for multiple time steps. While the predictions for Herwart look consistent in both datasets at first sight, fewer stations are available over Poland, where the CJ was overforecast, such that the overforecast in the gridded data carries more weight compared to the station data.

The reliability diagrams of the one-against-all approach for the COSMO-REA6 data (Figure 6.10) show that the calibration curves deviate more from the diagonal than for the station-based data (Figure 6.8) but are still reasonably close to calibrated. For the WJ and the CJ, we observe slight overforecasting (Figure 6.10b,d), whereas we observe underforecasting for the CFC (Figure 6.10c). For the CS, we observe a similar calibration curve to the station-based data (Figure 6.10e). The distinction of the individual features, which we assess via the all-pairs approach in Figure 6.11, results in mostly well-calibrated probabilities. The largest deviations from calibration are observed again for the distinction of the CJ and the CS, as discussed above, and for the distinction of the WJ and the CFC (Figure 6.11e), where the WJ is identified more frequently than observed. Overall, the predictions based on the COSMO-REA6 data are satisfactory considering that the RF models were trained on data from the station observations.

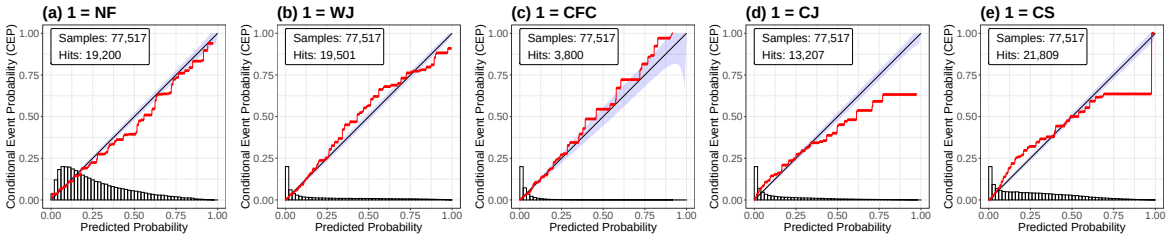

Figure 6.10: As in Figure 6.8 but based on COSMO-REA6 data.

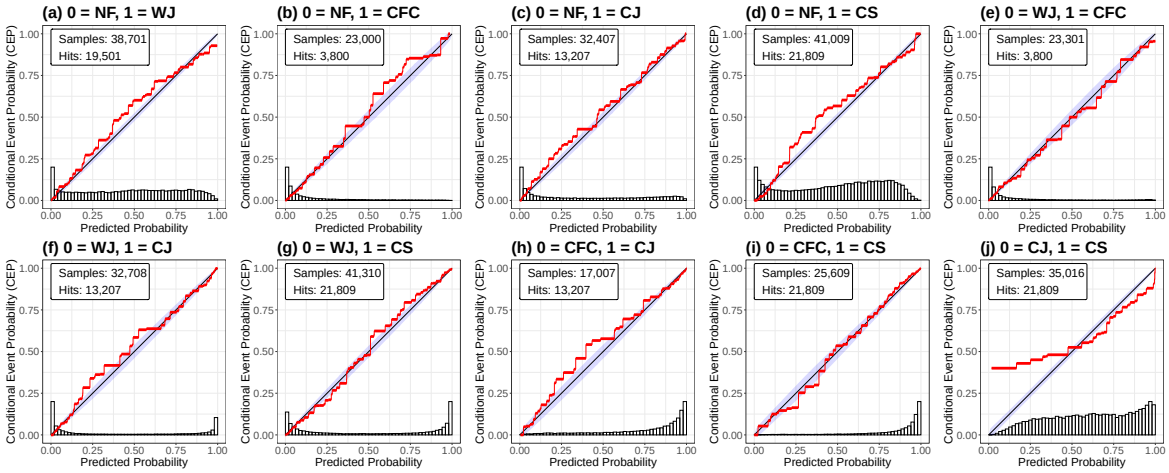

Figure 6.11: As in Figure 6.9 but based on COSMO-REA6 data.

PREDICTOR IMPORTANCE

To identify the predictors most relevant for the prediction of the wind features and the discrimination between two features, we calculate the BS permutation importance for the one-against-all and all-pairs approach. The BS permutation importance in the one-against-all approach is displayed in Figure 6.12. In general, Δp is the most important predictor variable, especially for the WJ. Only for CFC, it is not an important predictor, as it can occur slightly ahead of the cold frontal pressure trough, hence in a region of positive Δp. On the other hand, the absolute p values seem to be of less importance for WJ and NF, which occur further away from the cyclone center than CJ and CS, for which p indicates the proximity to the cyclone center. For CFC, we find instead that RR is the most relevant predictor variable as expected, while being less important for WJ, CJ and CS. For most features, d seems to be relevant, as it is a characteristic for the location relative to the cyclone center. This also leads to a high importance for NF occurring more frequently north or west from the cyclone center. However, d is not important for CFC, probably as convection leads to a more variable wind direction and due to the characteristic jump in d at cold fronts. To the contrary, Δd is of minor relevance for all features as well as $\Delta\tilde{\theta}$. A more important temperature-based predictor seems to be $\tilde{\theta}$, although again being less relevant for CFC. Lastly, $\tilde{v}$ shows its highest

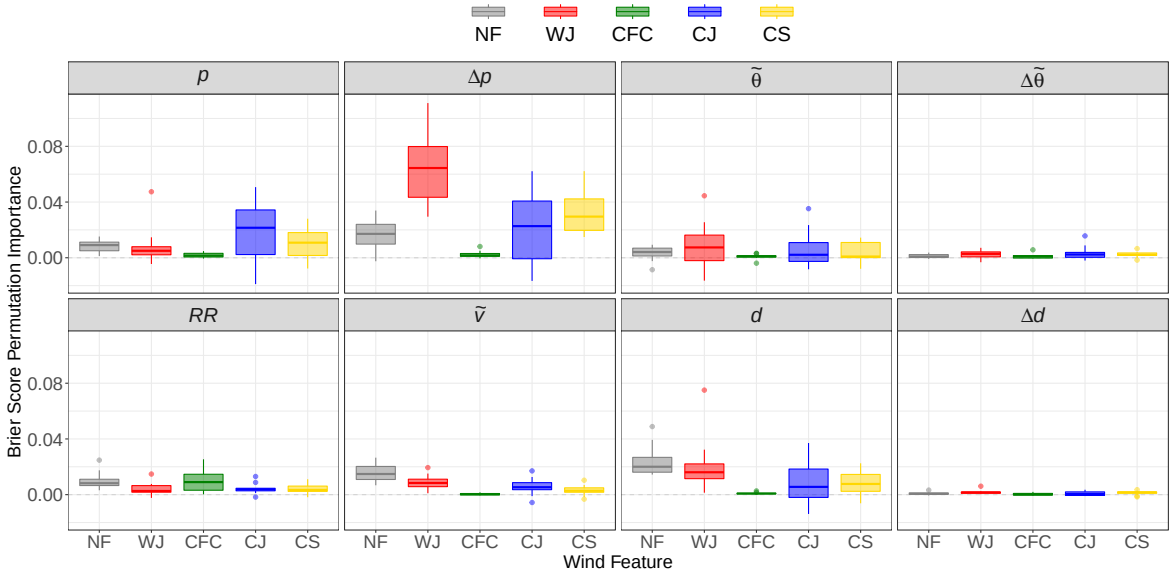

Figure 6.12: Boxplots of the BS permutation importance of the RF probabilities for the individual wind features and predictor variables in the one-against-all approach. The boxplots are calculated over the individual winter storms.

importance for NF, as higher wind speeds are less likely to be found at the boundary of a cyclone.

In the all-pairs approach (Figure 6.13), we can attribute the importance of the predictor variables more accurately. The key to distinguish the WJ from all other features is Δp, especially from the CJ and CS. This is consistent with the one-against-all discussion above. The large outlier in Δp in WJ vs. CJ is related to storm Herwart.[15] Of secondary importance is d, particularly when compared to CJ, CS and NF. Temperature also plays some smaller role in the distinction of the WJ. For CFC, the by far most important predictor is RR, but when compared against the CJ, p, Δp, $\tilde{\theta}$, $\Delta\tilde{\theta}$ and d also contribute. The positive outlier in RR is related to storm Fabienne (not shown). The distinction of the CJ to other features is more complex. As already discussed, p is relevant in all CJ-pairs. The distinction of CJ from NF additionally hinges upon Δp, $\tilde{\theta}$ and d.

The shortcomings of the RFs to distinguish CS and CJ are also reflected in Figure 6.13 by partly negative values for p, Δp and $\tilde{\theta}$. A negative value indicates that the RF probabilities perform better, when we break the link to the target variable by randomly permuting the predictor values. This is mostly due to storm Sabine, which reached an unusually low minimum core pressure of less than 950 hPa over the Norwegian Sea. Because of this, p values in the CS over continental Europe were similar to values typical of a CJ.

We do not only want to identify the most relevant predictors, but also investigate their effect on the predictions, which is illustrated for the eight predictor variables by the PDPs

[15] Discussed in Section 7 of Eisenstein et al. (2022).

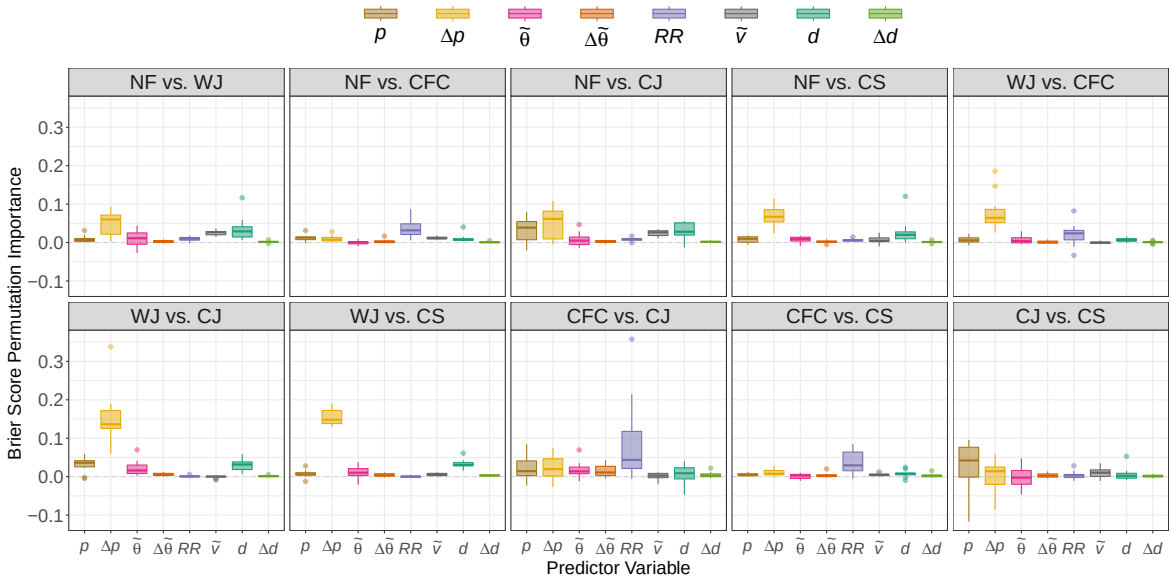

Figure 6.13: Boxplots of the BS permutation importance of the RF probabilities comparing two wind features for the predictor variables in the all-pairs approach. The boxplots are calculated over the individual winter storms.

in Figure 6.14. Again, the largest impact is found for Δp. The probability of observing a WJ is largest for small values of Δp and declines rapidly as the tendency increases and switches signs, while the probabilities of the CS and CJ increase. Probabilities for NF decrease slightly, while changes for CFC are small. For little RR, the probability of a CFC is close to zero, but consistently increases with increasing precipitation. In turn, probabilities for other features slightly decrease with increasing precipitation. In general, CJ and CS show high probabilities for low p values consistent with their occurrence during the most intense stage of a cyclone. However, surprisingly, CS shows higher probabilities than CJ between 970 to 980 hPa, although the CJ is usually closer to the cyclone center. This is again associated with the unusual behavior of storm Sabine, with its deep pressure minimum but no subjectively identified CJ. As such intense cyclones are rare, we are confident that the RF performs well in most more ordinary cases. As discussed previously, d is dependent on the location relative to the cyclone center. As the introduced features are all located south to west of the cyclone, we focus on values from 90° to 360° only. Within the WJ, d values mostly show south-westerly winds and do not change drastically. Probabilities for CFC increase with a positive wind shift, leading to more westerly and north-westerly winds for CFC but also following features, i.e., CJ and CS. Consistent to its low BS permutation importance, Δd shows almost no change in probabilities for all features. For $\tilde{\theta}$, an increasing trend for the WJ is shown, while the probabilities decrease for the other features, most strongly for the CJ, as one would expect. For $\Delta\tilde{\theta}$, we see indications of the air mass change at the cold front and thus higher probabilities in CFC for negative values. The CJ shows a slightly positive trend, while all the others are flat.

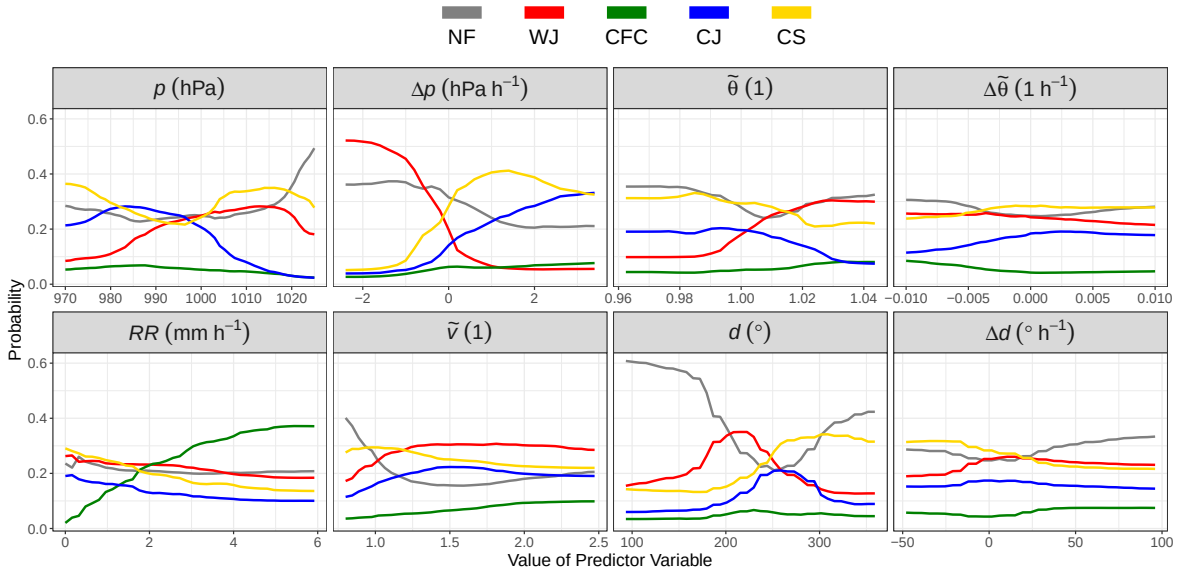

Figure 6.14: Partial dependence plots for the predictor variables and wind features.

Overall, investigating the importance of the predictor variables on the predictions, we find that the RFs largely learn physically consistent relations, as described in Section 6.3.2.

Discussion of spatial independence

The decision not to use spatial (nor temporal beyond 1 hour) dependencies in the identification algorithm makes our method highly flexible in its application, but the local approach can also cause issues where features deviate from their stereotypical characteristics. One example for the problem is the CJ of Sabine. Another example is storm Xavier, where for several hours many points within the vicinity of the cyclone show the highest probability for NF, rather than for any of the mesoscale wind features. The main reason for this appears to be that Xavier was characterized by unusually cool $\tilde{\theta}$ and high p (not shown), generally two of the most important parameters to distinguish features (Figure 6.12). While one predictor behaving in an unusual way could be compensated, e.g., in the case of Fabienne, two anomalous behaviors unsurprisingly result in considerably greater uncertainty.

A possible solution to the issues described here on the basis of Sabine, Xavier and Fabienne is to not only regard anomalies from diurnal and seasonal cycles but also to include some kind of spatial background, e.g., by normalizing p by the core pressure to detect the region close to the cyclone center, or comparing $\tilde{\theta}$ to the mean state over Europe during the period of the storm to detect the warm sector. However, such a step would bring its own set of problems. Any spatial mean would require an arbitrary decision about the considered area, which may vary greatly from cyclone to cyclone. Moreover, spatial means computed from surface observations are not representative due to the irregular spacing of the stations. Essentially, as

the features identified by the RF still occur in the expected areas, we conclude that a flexible local approach offers more advantages than shortcomings overall.

CONCLUSIONS

High wind and gust speeds can be caused by distinct mesoscale features within extratropical cyclones, which occur during different stages of the cyclone life cycle, in varying regions relative to the cyclone center and have distinctive meteorological characteristics (e.g., Hewson and Neu, 2015). These differences likely imply differences in hazardousness, forecast errors and, hence, risk to life and property.

To better understand, monitor and predict these mesoscale features, we developed RAMEFI, a first-ever objective identification method that is able to reliably distinguish the four most important features, that is, the WJ, the CJ, CFC and CS. The rare and often short-lived SJ is included in the CJ category, as their surface characteristics are often rather similar and 3D-trajectories are required for a clean distinction (Gray et al., 2021).

The first step was to subjectively label surface stations over Europe for 12 selected winter storm cases between 2015 and 2020. Based on the outcome, we trained a probabilistic RF based on the eight predictors $\tilde{v}$, p, Δp, $\tilde{\theta}$, $\Delta\tilde{\theta}$, RR, d and Δd. We note that we set a $\tilde{v}$ threshold of 0.8 to focus on high-wind areas. However, we do not expect the RF to be sensitive to small changes in the threshold and, in principle, the RF can be applied to wind speeds below this. Being independent of spatial behavior or gradients, the approach is very flexible and can be applied to single stations or grid points and various datasets with differing grid spacing, e.g., in the postprocessing settings of Chapter 5. However, due to the fast movement of meteorological features in stormy situations, hourly resolution is required, making the algorithm inapplicable to some climate datasets. To obtain areal information from irregular station data, Kriging was applied on the station-based probabilities generated by the RF.

The trained RFs are generally well-calibrated. Merely, the distinction between CJ and CS is more challenging, since the two features show similar characteristics in most parameters except for the fact that a lower value of p in the CJ is located nearer to the cyclone center. Overall, the RFs learn physically consistent relations reflected in the importance of individual predictors. For example, while Δp appears to be most important for WJ, CJ and CS, RR is substantial for the identification of CFC.

A detailed analysis of the RF feature probabilities for the selected cases shows a high consistency with the subjectively set labels with only few disagreements, mostly in cases of large deviations from standard cyclone models. While the identification of WJs has the highest confidence, the identification of CFC is least certain due to relatively few surface stations reporting hourly precipitation and thus less training data. Even the distinction between the relatively similar CS and CJ works well in most cases and time steps. In some cases, however, high probabilities of CJs are predicted by the RF in areas where no CJ was identified subjectively due to a missing hook-shaped structure and occlusion front, or too large

distance from the cyclone center (e.g., Herwart, Sabine). Despite the spatial independence of the method, putting the predicted probabilities together on a horizontal map and following the storm evolution in time shows a high degree of coherence for each feature (not shown), demonstrating the success of our method.

The station-based RFs are also applied to COSMO reanalysis data without any adaptations to the new dataset. Nevertheless, the obtained results are mostly consistent and only slightly less calibrated. This demonstrates that the method could be readily applicable to other analysis and forecast datasets, such as the COSMO forecast dataset used for wind gust prediction in Sections 5.3 and 6.1. Although applying RAMEFI over regions other than that used in the training has not been examined yet, its reliance on location-independent predictors suggests that it should be possible with no or only little modification.

Now that the RAMEFI method is fully developed, it enables a number of follow-on studies. One pathway is to use the objective identification approach to compute a long-term climatology over Europe based on station observations and COSMO reanalysis data. Although, previous literature discussed different causes of winds within extratropical cyclones, their climatologies were based on more subjective categorisations for a limited sample size (e.g., Hewson and Neu, 2015; Earl et al., 2017). RAMEFI will for the first time allow a statistically substantiated analysis of the characteristic of the mesoscale wind features in terms of size, lifetime, position relative to the cyclone core, occurrence relative to the lifecyle of the cyclone and wind characteristics. Furthermore, a systematic forecast error analysis can reveal to what extent forecast errors differ between the identified features and whether there are significant, systematic deficits in their representation in models. Here, however, we will not follow this approach for the final section of this chapter, but instead directly jump to the development of a feature-dependent postprocessing approach based on RAMEFI.

6.4 Experiments on feature-dependent postprocessing

Under the hypothesis that forecast errors are dependent on the high-wind features identified in Section 6.3, we want to develop a feature-based postprocessing approach that improves the predictive performance and ideally eliminates the forecast busts observed in Section 6.1. The foundation for feature-dependent postprocessing will be the RAMEFI method introduced in Section 6.3, as it is able to objectively identify the high-wind features independently of space and time (beyond 1 hour), based on a small subset of variables that can also be derived from forecast data.

The goal of the systematic comparison in Section 5.3 was to find the best postprocessing method for ensemble forecasts of wind gusts, with the result that NN-based methods outperformed all other approaches. Due to the comparable performance of DRN and BQN, we will focus on the conceptually simpler DRN in the following, as in Section 6.1. Note that we will not conduct a comprehensive analysis of a mature, fully developed method here, but instead

evaluate first ideas and experiments towards feature-dependent postprocessing that will make use of the NN-based postprocessing framework.

The final section of this chapter is structured as follows. First, we describe how the RAMEFI method is adapted towards the COSMO data underlying the case studies on wind gust prediction, with the goal to generate probabilistic predictions of the high-wind features. Then, we will present first approaches towards feature-dependent postprocessing incorporating the RAMEFI identification. At last, we will evaluate the performance of these approaches, both for high winds and winter storms in particular, and compare them with the standard DRN approach applied in Sections 5.3 and 6.1 to investigate whether an improvement was achieved by feature-dependent postprocessing.

6.4.1 High-wind feature forecasts via RAMEFI

To generate probability forecasts for the high-wind features, we apply RAMEFI to the COSMO-DE-EPS forecasts that have been used in Sections 5.3 and 6.1. While the application and statistical evaluation of RAMEFI in Section 6.3 had to rely on a CV setting, the RF underlying the high-wind feature forecasts can be trained on all case studies for this application. Analogous to the application of RAMEFI to the COSMO-REA6 data in the previous section, the RF is readily applicable for forecast data. However, not all of the predictor variables listed in Table 6.2 are output of the COSMO model, and need to be derived first.

Of the variables listed in Table 6.2, the wind speed, wind direction, mean sea level pressure, air temperature and precipitation are included as ensemble forecasts in the dataset and will be represented by the ensemble mean. Note that these variables coincide with those observed at the surface stations for training of the model (Table 6.2). As for the surface observations the RF was trained on, the potential temperature can be derived by first calculating the surface pressure p_s at altitude h via the barometric height formula, that is,

$$p_s = p\left(\frac{T}{T + 0.065h}\right)^{5.255}, \tag{6.5}$$

and then using the formula for the potential temperature (Lackmann, 2011, p. 9):

$$\theta = T\left(\frac{p_s}{10^5}\right)^{287.05/1005}. \tag{6.6}$$

In contrast to the RAMEFI predictor variables listed above, the calculation of the normalized wind speed and potential temperature requires climatological information in form of quantiles. For both variables, we normalize the underlying parameter either with the 98th percentile of the wind speed or the median of the potential temperature. Both quantiles are computed with respect to the location, time of the day and day of the year, as described in Section 6.3.

The three remaining predictor variables are tendency variables that are calculated as hourly increments. As the COSMO-DE-EPS dataset includes hourly lead times from 0 to 21 hours,

we can calculate the tendencies using the forecasts from the prior step for all but the 0 hour forecast. But, as discussed in Section 5.3, the 0 and 1 hour forecasts are subject to the spin-up effect and exhibit systematic deviations from the other lead times in terms of the bias and ensemble range. Due to these inconsistencies, we calculate the tendencies only for lead times of 3 hours or larger, and therefore exclude lead times affected by the spin-up effect.

Normalization of wind speed and potential temperature

Here, we investigate how to normalize the forecasts for wind speed and potential temperature in order to generate predictor variables for RAMEFI. To that end, we compare one approach based on station observations and one on COSMO forecasts.

First, we note that the normalization used in Section 6.3 is based on a 19-year climatology at the corresponding stations. Advantages of a station-based approach to normalize the COSMO forecasts are that it was used to generate the data the RF was trained on, and that the climatology is based on a large amount of observations. However, the major shortcoming is that the station data exhibits other characteristics than the model data, meaning that the quantiles are not subject to the model bias and do not accurately estimate the distribution of the model forecasts. Alternatively, we can generate a climatology based on the COSMO dataset, resulting in a consistent normalization. Based on the same arguments as for the postprocessing models in Chapter 5 and Section 6.1, we treat each initialization and lead time separately. Then, the size of the underlying dataset is smaller, since the COSMO data comprises only six years.

Further, we want to keep a clean split between training and test data, thus we exclude the test data from the climatology. In case of the model data, this means that we need to omit one year from the test data and ultimately each sample in the COSMO data comes with its own climatology.[16] However, for the station-based climatology, we neglect these concerns for a first analysis, as the effect of a single observation is smaller and we want to keep consistency with the RF. Altogether, the station-based climatology includes 399 samples (19 years times 21 days), while the model-based includes 105 (5 years times 21 days), if no values are missing (Table 6.5).

As the focus of RAMEFI is on European winter storms, we generate RF probabilities only for samples within the extended winter period from October to March, that is, the domain of the winter storms used for training RAMEFI (Table 6.3). While the quantile values used in Section 6.3 are readily applicable, those based on the model data need to be derived first. As mentioned above, wind speed is included in the dataset and potential temperature can be calculated via equations (6.5) and (6.6). Note that we use the model surface height (and not the station altitude) as the underlying surface height in equation (6.5).

Comparing the station- and model-based approach for the COSMO forecast data, we first

[16] The data selection scheme is identical to the CV approach used for training of the postprocessing models in Section 6.1.

Table 6.5: Overview of the approaches for normalizing wind speed and potential temperature. For a definition of the high-wind features, we refer to Section 6.3.

Approach for normalization	Surface stations		COSMO-DE-EPS	
Normalization				
Sample size for quantile estimation	399		105	
v: Number of quantile values (per location)	8,760		335,540	
θ: Number of quantile values (per location)	7,909		335,540	
Forecast data (yearly average, per lead time and initialization hour)				
Sample size	60,512		63,518	
Sample size during winter (half)	28,640		31,646	
High-wind samples during winter (% of winter)	1,741	(6.08)	2,692	(8.51)
Number of expected NF cases (% of high-wind)	873	(50.14)	1,378	(51.21)
Number of expected WJ cases (% of high-wind)	335	(19.22)	515	(19.15)
Number of expected CFC cases (% of high-wind)	162	(9.33)	257	(9.55)
Number of expected CJ cases (% of high-wind)	55	(3.13)	65	(2.43)
Number of expected CS cases (% of high-wind)	316	(18.17)	475	(17.66)

notice a major shortcoming of the observational approach, that is, the air pressure, which is required to calculate the potential temperature, is not reported at 17 of the 175 stations. Thus, the potential temperature forecasts cannot be normalized at these stations.[17] For the remaining stations, we compare the quantile values corresponding to each other. Note that we obtain different numbers of quantile values. As listed in Table 6.5, the observational dataset includes an average of 8,760 (7,909) values per location for wind speed (potential temperature). In case of wind speed, that is one for each hour of the year. As mentioned above, the median values of the potential temperature are not available for some stations. In contrast, the model data provides on average 335,540 values per location and variable, because we consider a CV approach for calculating the values and we treat each pair of lead and initialization time separately. Together, the approach based on the model data results in around 40 times more quantile values.

Figure 6.15 shows the differences of the quantile values dependent on the station, where negative values correspond to a larger estimated value for the station data. For the 98th percentile of the wind speed, we observe that for most of the stations the observational data has larger quantile values. Taking a look at the seven outlier stations with the largest deviations between the values, we find that these are stations located at a high altitude subject to larger systematic differences between model and station. For the median of the potential temperature, we find the same pattern with less outliers. Here, we observe that the difference is larger, the farther south the station is located.[18]

[17] For one of these 17 stations, the median is available at 0.5% hours of the year. Still, the station is excluded.
[18] Roughly, the stations are numbered from north to south.

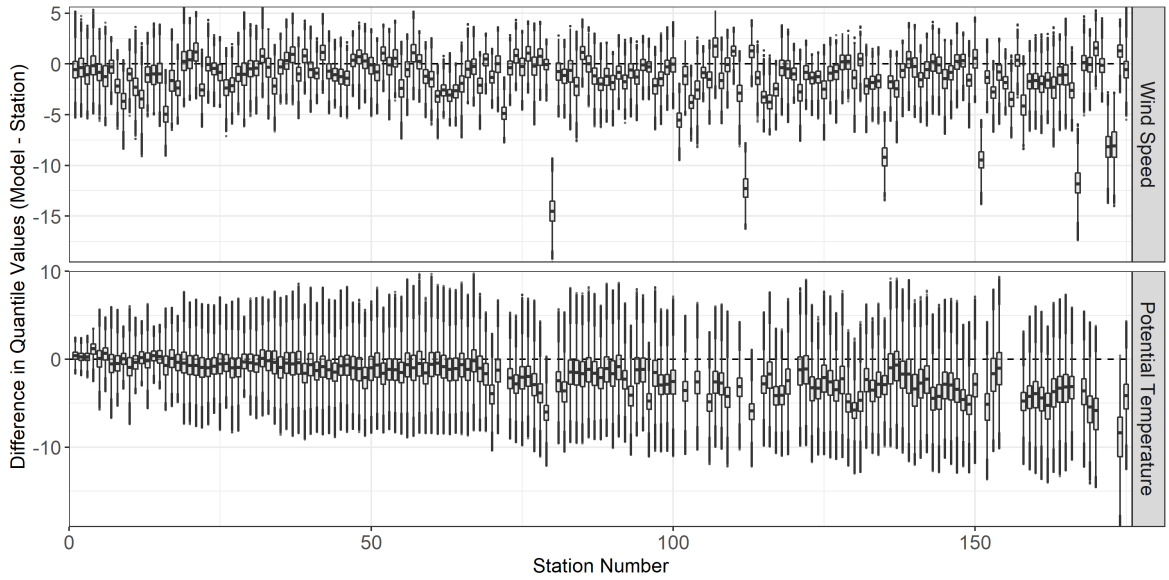

Figure 6.15: Boxplots of the differences in quantile values based on model and station data. Negative values indicate that the value is smaller for the model data.

The implications of the differences are described in the following. As RAMEFI is only applied when the normalized wind speed exceeds the threshold of 0.8, and as the normalized wind speed takes smaller values when a larger quantile is used, RAMEFI is applied to a smaller fraction of the data for the observational data. As winter storms, or high-winds in general, are extreme events, a larger amount of samples with threshold exceedance is preferred. Next to the classification of high-wind, we discuss the effect on the RAMEFI probabilities via the PDPs in Figure 6.14.[19] For the normalized wind speed, we find that larger values increase the probability of observing one of the high-wind features in general (instead of the no feature class NF). In case of the potential temperature, the smaller quantile values of the model data also result in larger values of the predictor variable. Based on the PDPs, this results, on average, in a shift from the colder features CJ and CS towards the WJ that is associated with warmer conditions.

For both of the approaches, we proceed by calculating the remaining predictor variables and then applying RAMEFI. In case of the model-based approach, the CV setting results in six different probabilities for each sample, specifically, one for each year that was left out. In contrast, only one probability is calculated for each sample in case of the observational data. Recall that we calculate the RAMEFI probabilities only for the period from October to March, otherwise we set them to zero.

Comparing the RAMEFI probabilities generated based on the two datasets, the average yearly number of samples with a threshold exceedance (for each pair of initialization hour and

[19]The PDPs in Figure 6.14 are based on the individual RFs of the CV approach, and not the RF used in this section. However, the differences are negligible.

lead time) is 1,741 based on the station data, that is, 6.08% of the winter period, while it is 2,692 for the model data, that is, 8.51% (Table 6.5). Hence, the differences due to using different quantile values are present in the RAMEFI identification.

Table 6.5 also lists the average number of expected high-wind features per year, which is calculated by accumulating the probabilities. Note that the RAMEFI probabilities in Section 6.3 were not only well-calibrated for surface observations but also for the COSMO-REA6 data that has the same underlying model as the COSMO-DE-EPS data. Only around half of the high-wind samples are classified as one of the four wind features, where the WJ and CS are most commonly predicted with 19 and 18%, respectively. Interestingly, CFC is predicted three times more often than the CJ. For the two approaches, the relative frequency of the features is fairly similar. However, we observe that NF is predicted more often for the station-based normalization, while CJ and CS are predicted less often. This coincides with what we expected when comparing the quantile values.

Altogether, we decide to use only the probabilities generated based on the model data. The main reasons are that, due to missing values and smaller high-wind thresholds, more samples are available for the model-based approach (Table 6.5), and that the station-based approach is not sufficiently consistent with the forecast data (Figure 6.15). Therefore, more RAMEFI probabilities are generated and larger subsets relevant for feature-dependent postprocessing are available.

6.4.2 Feature-dependent postprocessing models

In general, regime-dependent postprocessing approaches provide promising pathways for future work (Rodwell et al., 2018), and they have demonstrated their potential in different studies (e.g., Gneiting et al., 2006; Allen et al., 2021). Here, we want to highlight one aspect connected to the choice of the underlying weather regimes. The finer the distinction between the meteorological conditions, the closer we can adapt to the specific situation at hand. However, the more regimes are chosen, the smaller the amount of data associated with the regime. This is the same trade-off inherent in training data selection processes, e.g., when choosing between local and global training as described in Section 5.1. Many approaches to regime-dependent postprocessing operate by training separate models for the different weather regimes, and then generating forecasts using the distinct model instances (Allen et al., 2020, 2021). Based on probabilistic predictions of the regime occurrence, this can be done by selecting the most likely or combining several models based on their likelihood, among other variants.

In this section, the regimes are given by the high-wind features, hence we refer to feature- instead of regime-dependent postprocessing. This definition involves two nested distinctions, in particular, that of high-winds and that between the features, but only in case of high-winds. Table 6.5 shows the average occurrence of high-wind events and the expected number of features, using a model-based normalization. The ratio of high-wind events is only at about

8.5% of the samples during the winter period, and the individual features are observed much less often. Still, note that a particular advantage of a probabilistic approach for feature identification is that we obtain occurrence probabilities for all high-wind samples. As discussed in Section 6.3, given a binary identification, no information on the likelihood of the feature, hence, joint occurrences, transition zones or ambiguous situations, is available.

We choose to focus on feature-dependent postprocessing models that explore the flexibility of the underlying NN structure via the ideas of *transfer learning* (TL), in addition to reference approaches for comparison. Recall that we presented a general NN-based postprocessing framework in Section 4.3, hence the following approaches can also be directly applied for BQN and HEN.

Reference approaches

Next to the standard DRN approach, which we will refer to as REF, two other reference approaches for feature-dependent postprocessing are implemented. The RFP approach simply uses the RAMEFI probabilities as additional predictor variables for DRN. The probabilities are processed analogously to the other predictor variables, which includes a normalization, and the hyperparameters are not adapted. As the ratio of high-wind events in the training data is small, we expect the RAMEFI probabilities not to have an effect on the predictions, and not to be considered important by the NNs. The intention is to show that it is not sufficient to supply only the identification to the model, instead of adapting the model itself.

The second approach is not based on the wind features but instead on the classification of high-winds; we simply apply DRN separately for normal and high winds. As we fit two DRN models, this approach is referred to as 2-REF. Here, the DRN model for the high-wind data has substantially less data available, which might deteriorate the predictive performance. As for all of the reference approaches, we do not adapt or tune the hyperparameters of the standard DRN approach. The 2-RFP model combines the RFP and 2-REF model, as it estimates separate DRN models for normal and high winds, but includes the RAMEFI probabilities as additional predictors (for the high-wind model). The intention behind 2-REF and 2-RFP is to investigate whether a separate NN model for the high-wind data is sufficient, or if more sophisticated approaches are required.

Transfer learning

We choose to focus on feature-dependent postprocessing approaches tailored to the NN-based framework based on the ideas of TL (e.g., Pan and Yang, 2010; Goodfellow et al., 2016). As the name suggests, the general concept of this approach is to transfer information from the application on one domain to another. More specifically, this includes to adapt NNs trained on large datasets that have proven to perform well in the application at hand (such as DRN in Section 5.3) towards another familiar application by modifying the architecture, retraining

the existing parameters and fine-tuning the network. TL approaches are typically used when it is not feasible to build an own, task-specific NN from scratch, e.g., if a sufficient amount of data is not available or well-performing NNs from related applications provide promising base models. Further, methods from TL are suitable in situations where the underlying data generating mechanism changes, e.g., a far-reaching update of the NWP model such as the *kilometer-scale ensemble data assimilation system* (KENDA; Schraff et al., 2016) in case of the COSMO-DE-EPS (Pantillon et al., 2018, Section 2.1), or where a focus lies on extreme events, such as in this chapter. For reviews on TL, we refer to Weiss et al. (2016) and Zhuang et al. (2021).

Here, we will apply TL for feature-dependent postprocessing proceeding in three steps. First, a DRN model is fitted on the entire data, as is standard. This base model will be used to predict non high-wind samples, while a second model will be estimated for high-wind data. This second model extends the standard DRN architecture by incorporating the RAMEFI identification. We will present two possible architecture choices to include the probabilities as additional predictor variables, but also one reference model not based on RAMEFI, with the intention to filter out the effect of using TL without additional information. For training of the extended model, we want to make use of the existing model as much as possible. Therefore, the second step of our TL approach involves the transfer of the NN parameters, i.e., the weights and biases, from the base model to the extended one. As the extended model includes new and omits old connections, not all existing parameters can be transferred to the new model. The second step is then to train (only) the new connections of the extended model on the high-wind data, i.e., we freeze the part of the model that was transferred from the base model by setting the learning rate to 0 for this part. In this step, we slightly adapt the (overall) learning rate, number of epochs and batch size. The third and last step is typically referred to as fine-tuning. Including all parameters in the estimation process, i.e., unfreezing those from the base model, we fit the model again on the high-wind data.

The aforementioned reference model does not extend the DRN architecture, but simply includes retraining on the high-wind data not making use of the RAMEFI identification, and will be referred to as TL-REF. Hence, the second step of the process can be omitted. The two more interesting approaches involve an extended model that incorporates the RAMEFI probabilities, which is illustrated in Figure 6.16 for both the TL-1 and TL-2 approach.

The extended architecture of the TL-1 approach connects the RAMEFI identification with the first layer of the network, i.e., the feature probabilities are treated analogously to the other predictor variables (including the normalization described in Section 5.3.2). For the initial estimation of the high-wind network, only the connections from the input nodes towards the first hidden layer are trainable. The parameters from the input layer to the first hidden layer that have been estimated in the base model, i.e., all connections not involving the RAMEFI probabilities, are used as initial values in the high-wind model.

The idea behind the TL-2 approach is to build an architecture that includes a channel for

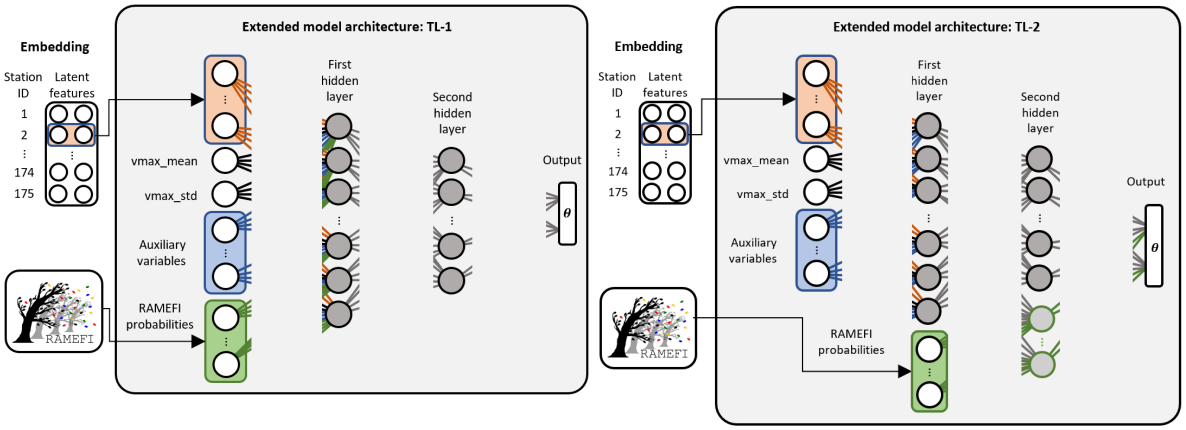

Figure 6.16: Graphical illustration of the extended model architectures applied in the TL-1 (left) and TL-2 (right) approach.

the feature information parallel to the base model. This is achieved by connecting the input nodes of the RAMEFI probabilities only with a separate set of nodes in the second hidden layer, which are also connected to the first hidden layer of the base model. This leaves the base model untouched and combines the information extracted in the first hidden layer with the feature information. The second hidden layer is then connected to the output nodes as usual. In the initial estimate, the parameters of the new nodes in the second layer are, as well as those of the output layer, trainable. Here, we use the existing parameters for the output layer as initial values.

For both TL-1 and TL-2, we adapt the hyperparameters for training of the high-wind model by reducing the learning rate to 10^{-5}, decreasing the number of epochs to 50 and decreasing the batch size to 32. These values have not been determined using a strategy for hyperparameter tuning but rather on general suggestions for TL models and impressions from supervising the training of the networks. The number of nodes connected to the RAMEFI probabilities in the second layer of the TL-2 approach is 8. For a comparison with the standard DRN, we refer to Table 5.10.

Before assessing the predictive performance of the feature-dependent postprocessing approaches, we want to note that these models are only first approaches towards a feature-dependent postprocessing. A systematic comparison of different strategies for data preparation, possible NN architectures or hyperparameter choices may lead to other configurations as those presented in this section. Hence, there may be room for (substantial) improvement of these models.

6.4.3 Results

Replicating the case studies on wind gust prediction, we will evaluate the predictive performance of the feature-dependent postprocessing methods presented in the previous section.

Analogous to the systematic comparison in Section 5.3, we compare the postprocessing methods for the entire year of 2016. However, as (basically) all models rely on the standard DRN for the non high-wind data, that is, over 95% of the data, we will consider only the high-wind samples for a more detailed investigation.[20] A quick look at the CRPSS including all samples confirms that all methods perform similar (Figure 6.17). Recall that the definition of high-winds is based solely on the ensemble predictions, hence we do not face undesired effects from conditioning on extreme events in the evaluation (Lerch et al., 2017). After covering the high-winds in 2016, we will investigate whether the approaches are able to eliminate the forecast busts in the application on winter storms (Section 6.1).

Comparison for high winds

To evaluate the effect of the feature-dependent postprocessing, we assess the predictive performance only for the high-wind samples of the data in Section 5.3, which results in test sets of average size 2,187. That is even less than the average size listed in Table 6.5. The standard evaluation metrics we also considered in the previous case studies are shown in Figure 6.17 for the feature-dependent postprocessing approaches.

The first observation is that the 2-REF and 2-RFP models perform constantly worse than the other approaches with a gap in the CRPSS of around 2%. These two approaches also predict higher wind speeds in general, as the bias of the median forecasts shows. This may be a result of the fact that this model is developed exclusively using high-wind data, unlike the other approaches. Further, the reason for the inferior predictive performance is the insufficient amount of training data. Hence, it is not sufficient to build a separate NN model for high winds. Due to the shortcomings of the purely threshold-dependent models 2-REF and 2-RFP, we will omit them from the following analysis.

The skill of the other approaches is on a similar level. The two reference models REF and RFP perform equally well, followed by the TL approaches in the order TL-REF, TL-2 and TL-1. While consistent over lead time, the gap between the models is considerably small. Comparing REF and RFP, we do not find any systematic differences between the two models. As supposed in the previous section, the inclusion of the RAMEFI probabilities as additional predictor variables without any adjustment to the structure of the model does not affect the predictive performance at all. A systematic difference in the comparison of the two reference models REF and RFP with the TL approaches is that TL results in sharper forecasts, here, exemplified by the shorter PIs. Comparing the PI coverage of the two groups, we find that the smaller PI lengths come at the cost of calibration. Overall, this results in less skill, as noted before.

[20] Note that the DRN (base) models underlying the different approaches are not identical due to practical reasons. Further, note that the RAMEFI probabilities are set to zero in case of non high-wind data and that 2-REF and 2-RFP do not use the entire training set to fit the standard model but only the non high-wind data.

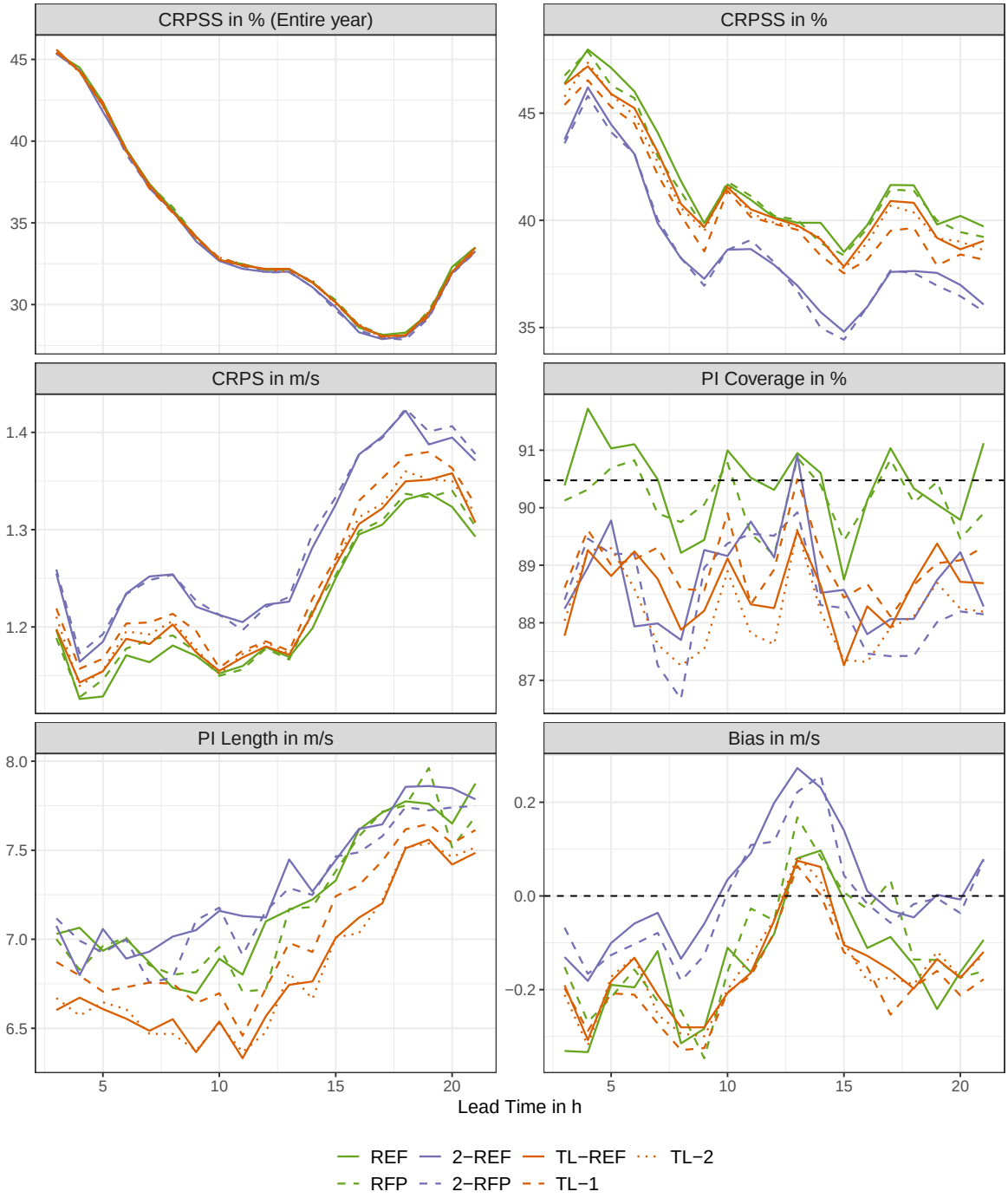

Figure 6.17: CRPSS with respect to the raw ensemble predictions, mean CRPS, mean PI coverage, mean PI length and mean bias of the feature-based postprocessing approaches as functions of the lead time, averaged over all stations for the entire year 2016 (topleft) and for samples that exceed the high-wind threshold (otherwise).

Within the TL approaches, we observe that the skill is largest for the TL-REF approach that does not extend the model architecture. Still, retraining without the inclusion of RAMEFI results in sharper PIs than the TL-1 approach that uses the feature probabilities as additional predictors in the first layer. The TL-1 approach also performs worse than TL-2, as the CRPS is consistently smaller. This might hint at the fact that the RAMEFI probabilities should be included customized at a later stage in the network such that the prevalent information does not get lost among the large number of other predictor variables.

Comparison in winter storms

Replicating Section 6.1, we apply the same CV setting for the feature-dependent postprocessing methods. First, we take a look at Figure 6.18 that extends the original panel in Figure 6.1 by the feature-dependent approaches. The main message is that the new methods were not able to eliminate the forecast busts. Instead, we still observe them and, in case of Joachim, even to a larger extent, especially for the TL approaches. Looking at the individual storms, we find that all postprocessing methods behave similarly in terms of bias, PI length and coverage, e.g., in terms of the bias for Andrea or the PI coverage for Christian. Extending Figures 6.3 and 6.4 that compare the methods for the individual storms, we found no noteworthy differences from the general behavior (not shown). Recall that for each time step and storm, the evaluation is based on at most 175 stations and that the underlying DRN model instances are not identical.

In Figure 6.19, we compare the predictive performance averaged over all storms to draw general conclusions in case of the winter storms (analogous to Figure 6.2). We note that the general behavior is similar to that observed for high-winds in 2016. The TL approaches perform slightly worse than the two reference approaches REF and RFP in terms of the CRPS. Again, the PI length is shorter and the coverage smaller for the TL approaches. Further, the TL approaches are slightly more biased towards smaller wind speeds. Within the groups, we find no differences.

Conclusions

In this section, we presented first models towards feature-dependent postprocessing based on the RAMEFI identification of Section 6.3, including both reference approaches and more sophisticated TL models. A comparison of these methods with the standard DRN in the setting of Sections 5.3 and 6.1 shows that none of the feature-dependent postprocessing methods did improve the predictive performance, neither for high-winds in general nor for selected winter storms, instead they performed slightly worse than the benchmark. Although the TL approaches result in sharper PIs, the coverage deviates more from the nominal level, therefore we have less reliable forecasts. Within the TL methods, the least complex variant closest to the standard approach (TL-REF) is preferable. Hence, the feature-dependent postprocessing methods introduced in this section are not able to improve predictive performance and

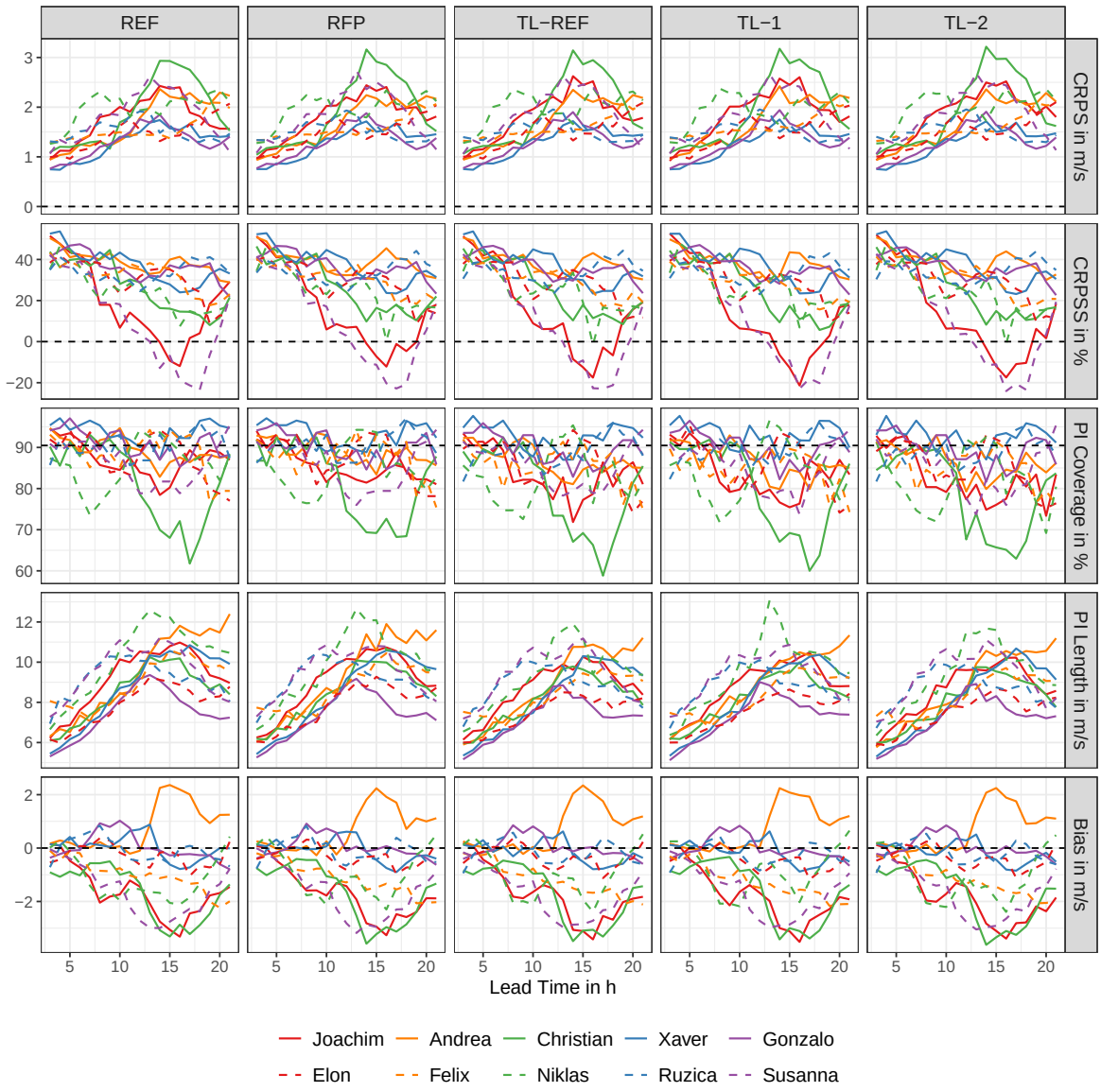

Figure 6.18: Mean CRPS, CRPSS with respect to the raw ensemble predictions, mean PI coverage, mean PI length and mean bias of the feature-based postprocessing approaches as functions of the lead time, averaged over all stations for each winter storm.

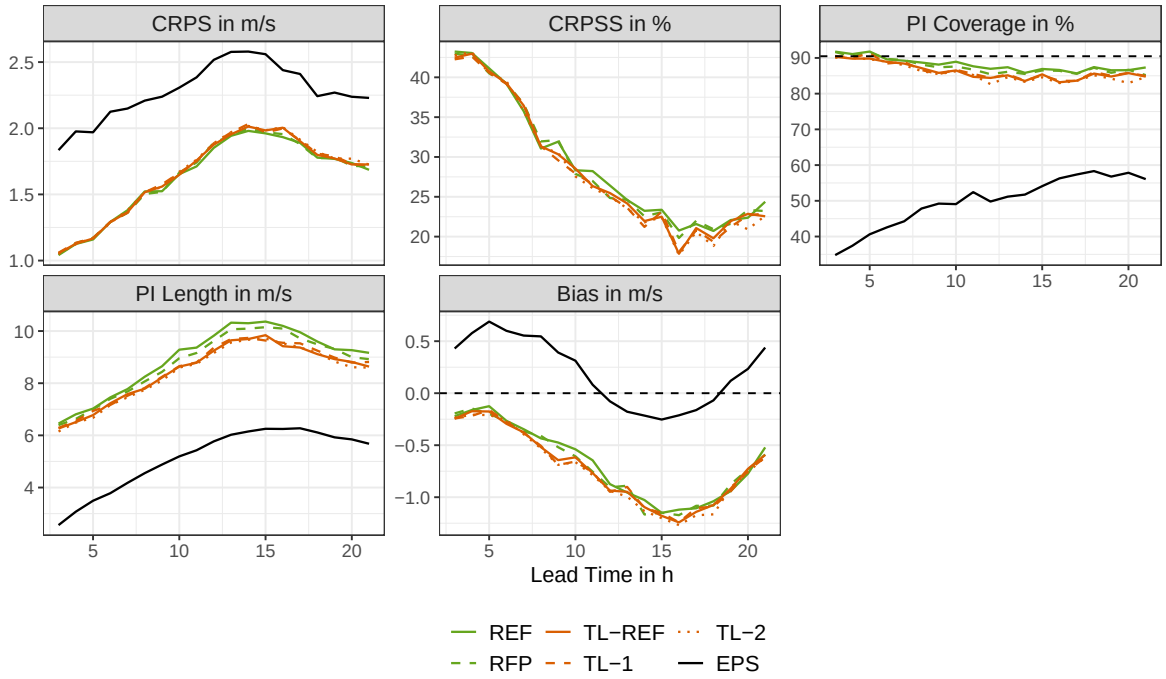

Figure 6.19: Mean CRPS, CRPSS with respect to the raw ensemble predictions, mean PI coverage, mean PI length and mean bias of the feature-based postprocessing approaches as functions of the lead time, averaged over all stations and winter storms.

eliminate the forecast busts observed for the winter storms. The inferiority of the newly introduced approaches has several possible explanations:

- First, the feature-dependent models presented in this section are not fine-tuned, in contrast to the DRN benchmark. A thorough investigation of optimal choices for the model architecture, hyperparameters, incorporation of the RAMEFI identification and data selection for training and validation, may improve the predictive performance. However, the small amount of high-wind data and features identified (Table 6.5) complicates the matter, specific measures such as a CV approach may be required to avoid overfitting. Further, the amount of data available may be insufficient to correct for (potential) feature-dependent errors.

- In this section, we operate under the hypothesis that forecast errors are dependent on the high-wind features. Thus, a feature-dependent error analysis is required to analyze more deeply whether and if so how forecast errors depend on the features. Further, this analysis can be used to adapt the methods towards the observed dependencies, e.g., errors might systematically deviate only for selected features.

- Recall that the predictor variables are based on the ensemble mean, which blurs the information present in the individual members. One example of an effect of using the

ensemble mean might be the higher number of CFCs identified by RAMEFI, since small and sharp precipitation fields at different locations in the individual members result in a large area for which precipitation is predicted after averaging. Note that precipitation is the most important predictor variable for identifying CFC (Figures 6.12–6.14). Hence, the application of RAMEFI on the individual ensemble members should be considered, as RAMEFI takes physical consistencies into account. In addition, the sample size could be increased drastically by using the full ensemble instead of the mean predictions.

– The reason for observing forecast busts for selected winter storms may lie at the storm-scale, i.e., the entire cyclone is predicted too weak, strong or shifted in the COSMO-DE-EPS forecasts. If the model predictions are subject to such systematic errors, the RAMEFI identification will not be reliable, as it does not correct for these errors. Recall that RAMEFI was trained and evaluated based on station observations and reanalysis data, which are not subject to misplacement errors at the storm-scale. Analogous to that in Section 6.3, a separate evaluation of the RAMEFI identification based on the COSMO forecasts needs to be conducted to assess the predictive performance and see whether the probabilities are (to a sufficient degree) accurate and reliable. Again, averaging over individual storm tracks in the ensemble members might be counterproductive.

– At last, there is a large case-to-case variability for the winter storms. In practice, the features seldom show textbook-like behavior and exhibit different characteristics, e.g., due to the influence of the local orography like land, sea or mountains. With this in mind and recalling that RAMEFI was trained on subjective labels of the wind features, the information present in the RAMEFI identification based on the model forecasts might simply not be sufficient to generate confident predictions that allow for a feature-dependent postprocessing approach.

Altogether, we conclude that all of these points need to be addressed to investigate whether a feature-dependent postprocessing approach is actually able to improve the predictive performance significantly and eliminate the forecast busts within the winter storms.

CHAPTER 7

Conclusions

In this thesis, we have demonstrated how methods from modern machine learning can be used to leverage the predictive performance of statistically postprocessed weather forecasts. In a systematic comparison, we found that the best postprocessing methods for ensemble forecasts of wind gusts are neural network-based approaches, due to their ability to incorporate additional meteorological predictor variables and spatio-temporal information. Beforehand, we formulated concrete methodological recommendations on the aggregation of deep ensembles, such as for the application on postprocessing. With the goal of eliminating forecast busts within winter storms by incorporating domain knowledge in neural network-based postprocessing, we developed a calibrated random forest-based identification of high-wind features, and demonstrated that it learned how an experienced meteorologists would complete that task. This concluding chapter aims to summarize the findings in this work, and addresses possible directions for future work.

In Chapter 3, we conducted a systematic analysis of aggregation methods for distributional forecasts generated by randomly initialized deep ensembles, i.e., ensembles of neural networks. Firstly, our findings coincide with the fact that forecast combination in general and, in particular, ensembling of machine learning methods improves predictive performance. We compared the two distinct approaches of averaging probabilities, referred to as linear pooling, and Vincentization, that is, averaging quantiles, which proved to be more suitable in our simulation and case studies. While the standard approaches to quantile and probability averaging do (in general) not correct for systematic errors present in the network predictions, we demonstrated that a general Vincentization framework is able to correct for biases and dispersion errors, and can be integrated in the typical training scheme of a neural network. Further, we found that the optimal ensemble size in case of the randomly initialized deep ensembles is approximately given by 10. However, deep ensembles can not only be generated based on stochastic gradient descent methods and random initialization but also using a wide range of alternative approaches such as Monte Carlo dropout (Srivastava et al., 2014; Gal and

Ghahramani, 2016), bagging (Breiman, 1996), Bayesian neural networks (Neal, 2012; Jospin et al., 2022), generative models (Mohamed and Lakshminarayanan, 2016) or BatchEnsemble (Wen et al., 2020). An interesting avenue for future work is to investigate whether the conclusions drawn in this study hold for deep ensembles based on these approaches and how they compare against each other. Even further, aggregation methods that simultaneously recalibrate the deep ensemble (members), such as the general Vincentization framework or probability-based calibration methods such as the beta-transformed linear pool (Ranjan and Gneiting, 2010), could be integrated in the estimation process of the network (e.g., Kim et al., 2021).

Chapter 4 presents a wide range of statistical postprocessing methods that can be divided in three groups of increasing complexity. In Chapter 5, we apply these methods in three case studies. The first group of methods consists of simple approaches rooted in statistics that are based only on the variable of interest. As demonstrated in all case studies of Chapter 5, these basic methods are able to significantly improve the ensemble predictions and yield well-calibrated predictions. In a pseudo-operational setting, the ensemble model output statistics approach could straightforwardly be adapted to a wide range of meteorological variables and implemented for near real-time postprocessing on the KIT-Weather portal. Another advantage of these approaches is that only small amounts of data are required for a successful application, as demonstrated for solar irradiance forecasts over Hungary. Still, these approaches are inherently limited by their parsimony and are inferior to more sophisticated postprocessing methods, as demonstrated in the case study on probabilistic wind gust prediction. While the second group of methods consists of established machine learning approaches based on the ideas of gradient boosting and random forests, the third is based on a common neural network framework. Both significantly outperform the postprocessing benchmarks of the first group as they are able to incorporate additional predictor variables and model more complex, nonlinear relations. While all postprocessing methods yield calibrated forecasts, superior predictive performance is achieved by maximizing the sharpness, following the central paradigm of probabilistic forecasting. Within the advanced methods, neural network-based postprocessing outperforms the established machine learning methods, as station embeddings allow to build a locally adaptive network model that can be trained on the entire training set.

The ability to adapt neural networks to the situation at hand, here exemplified by the use of a simple station embedding technique, provides manifold avenues for future research, as numerical weather prediction (NWP) models generate large datasets with intervariable and spatio-temporal dependencies, which we did not explore in this thesis. Convolutional or recurrent neural networks, among other variants, can be used to extend our postprocessing framework to include spatial fields or time series data as input variables (e.g., Gasthaus et al., 2019; Scheuerer et al., 2020; Veldkamp et al., 2021). On the other end, multivariate forecasts can be generated using, e.g., generative models (Chen et al., 2022), generative adversarial networks (Dai and Hemri, 2021), or alternatively by implementing neural network-based

member-by-member postprocessing, which preserves the multivariate dependencies in the NWP predictions leading to more realistic and physically consistent forecasts that are less prone to forecast busts.

In Chapter 6, we explored the idea of developing hybrid models that incorporate domain knowledge into the statistical postprocessing models. After identifying the best postprocessing method for wind gusts in Chapter 5, we investigated the predictive performance within European winter storms motivated by a study of Pantillon et al. (2018). Applying neural network-based postprocessing, we find that even sophisticated models are subject to forecast busts. Due to the occurrence of specific high-wind features, Pantillon et al. (2018) formulate the hypothesis that the forecast busts are associated with certain meteorological conditions. Hence, the incorporation of domain knowledge in the statistical postprocessing models, resulting in a hybrid model, may be able to eliminate such forecast busts. Before that, we demonstrate the potential of such hybrid models in a brief excursion towards tropical cyclone forecasting in the North Atlantic, where a hybrid model outperforms benchmark models by combining NWP forecasts and climatological information. Returning back to European winter storms, we developed RAMEFI, an automatic, objective identification that generates probabilistic predictions of high-wind features. Based on RAMEFI, we conducted first experiments towards feature-dependent postprocessing using ideas of transfer learning.

As aforementioned approaches were not able to eliminate the forecast busts, more work is required to reach this goal. Although we found that the networks in Section 5.3 and the random forests in Section 6.3 learn physically consistent relations between predictor and target variables, we were not able to identify why the networks adapted their predictions in the European winter storms. Here, more sophisticated methods for interpretable machine learning are required to explain the behavior of the neural networks (e.g., Molnar, 2018; McGovern et al., 2019). Especially for high-impact events, trustworthiness is an essential ingredient for forecasting. One step that we skipped in the previous chapter was to check how forecast errors depend on the high-wind features. Based on a feature-dependent error analysis, one can design a hybrid model that specifically addresses the shortcomings of NWP predictions for the different wind features. The feature-dependent postprocessing approaches presented in this thesis are spatially and temporally (beyond 1 hour) independent, however, the high-wind areas are coherent in time and space. This can not only be taken into account for model development but also forecast verification. Using common univariate verification metrics, a displaced high-wind area is subject to a double penalty. Hence, the development of mathematically principled spatial verification tools is another interesting avenue for future work (e.g., Brown et al., 2011; Skok and Hladnik, 2018).

Although the postprocessing methods are in principle readily applicable for operational use, there are many obstacles in the transfer from research to operations, e.g., preservation of physical consistency, adaptability towards model changes or technical challenges in the implementation (e.g., Taillardat and Mestre, 2020; Vannitsem et al., 2021). While postpro-

cessing methods presented in academic studies are often tailored to a concise problem at hand, weather services are in general interested in a common, robust postprocessing framework applicable to a wide range of forecasting applications rather than a multitude of complex, individual solutions. Regarding the methods presented in this work, one aspect for operational use would be to take measures in order to avoid forecast busts such as those observed for the near real-time postprocessing, as forecast busts are a serious concern for weather services. Further, in order to make (optimal) use of the postprocessed forecasts supplied, operational forecasters need to be aware of the advantages and shortcomings of the postprocessed products, e.g., via training. Especially methods from explainable artificial intelligence (AI) are a key tool for overcoming skepticism towards AI methods widely considered as black boxes. Apart from the use at weather services, the postprocessing methods presented in this thesis provide an excellent starting point for forecasting (sources of) renewable energy in the energy sector (e.g., Phipps et al., 2022; Gneiting et al., 2023).

Parallel to applications for statistical postprocessing, AI methods have made rapid progress for purely data-driven forecasting, challenging the prevalent practice of physics-based NWP models with the ultimate goal of replacing them (e.g., Schultz et al., 2021; Bi et al., 2022; Keisler, 2022; Lam et al., 2022; Pathak et al., 2022). While we suppose that NWP will not become obsolete and remain an integral part of weather forecasting in the foreseeable future, the success and potential of AI leads to the question whether the need for postprocessing will persist as these methods are surpassing NWP. Even under the bold prediction that AI models replace NWP, the forecasts will (most certainly) still be subject to systematic errors, or at least improvable, when evaluated for a particular objective. Recall that postprocessing allows to customize forecasts with respect to a specific target variable, such as the wind speed observed at a given site or the occurrence of a tropical cyclone. Hence, statistical postprocessing will still be needed to provide optimized predictions tailored to the problem at hand. To that end, transfer learning might become a key tool for statistical postprocessing of AI-based weather prediction models. Instead, the principles underlying the models developed in this thesis are universally applicable in the context of probabilistic forecasting and therefore provide valuable tools to leverage AI-based weather prediction.

Bibliography

Aastveit, K., Mitchell, J., Ravazzolo, F., and van Dijk, H. K. (2018). "The evolution of forecast density combinations in economics". *SSRN Electronic Journal*.

Abe, T., Buchanan, E. K., Pleiss, G., Zemel, R., and Cunningham, J. P. (2022). *Deep ensembles work, but are they necessary?* Preprint, available at `https://arxiv.org/abs/2202.06985v1`.

Akaike, H. (1974). "A new look at the statistical model identification". *IEEE Transactions on Automatic Control*, 19 (6), 716–723.

Alessandrini, S., Delle Monache, L., Sperati, S., and Cervone, G. (2015). "An analog ensemble for short-term probabilistic solar power forecast". *Applied Energy*, 157, 95–110.

Allaire, J. and Chollet, F. (2020). *keras: R interface to 'Keras'*. R package version 2.3.0.0. `https://cran.r-project.org/package=keras`.

Allaire, J. and Tang, Y. (2020). *tensorflow: R interface to 'TensorFlow'*. R package version 2.2.0. `https://cran.r-project.org/package=tensorflow`.

Allard, D., Comunian, A., and Renard, P. (2012). "Probability aggregation methods in geoscience". *Mathematical Geosciences*, 44 (5), 545–581.

Allen, S., Evans, G. R., Buchanan, P., and Kwasniok, F. (2021). "Incorporating the North Atlantic Oscillation into the post-processing of MOGREPS-G wind speed forecasts". *Quarterly Journal of the Royal Meteorological Society*, 147 (735), 1403–1418.

Allen, S., Ferro, C. A., and Kwasniok, F. (2020). "Recalibrating wind-speed forecasts using regime-dependent ensemble model output statistics". *Quarterly Journal of the Royal Meteorological Society*, 146 (731), 2576–2596.

Bakker, K., Whan, K., Knap, W., and Schmeits, M. (2019). "Comparison of statistical post-processing methods for probabilistic NWP forecasts of solar radiation". *Solar Energy*, 191, 138–150.

Baldauf, M., Seifert, A., Förstner, J., Majewski, D., Raschendorfer, M., and Reinhardt, T. (2011). "Operational convective-scale numerical weather prediction with the COSMO model: Description and sensitivities". *Monthly Weather Review*, 139 (12), 3887–3905.

Baran, Á., Lerch, S., El Ayari, M., and Baran, S. (2021a). "Machine learning for total cloud cover prediction". *Neural Computing and Applications*, 33 (7), 2605–2620.

Baran, S. and Baran, Á. (2021). *Calibration of wind speed ensemble forecasts for power generation*. Preprint, available at `https://arxiv.org/abs/2104.14910`.

Baran, S. and Lerch, S. (2015). "Log-normal distribution based ensemble model output statistics models for probabilistic wind-speed forecasting". *Quarterly Journal of the Royal Meteorological Society*, 141 (691), 2289–2299.

Baran, S. and Lerch, S. (2016). "Mixture EMOS model for calibrating ensemble forecasts of wind speed". *Environmetrics*, 27 (2), 116–130.

Baran, S. and Lerch, S. (2018). "Combining predictive distributions for the statistical post-processing of ensemble forecasts". *International Journal of Forecasting*, 34 (3), 477–496.

Baran, S. and Nemoda, D. (2016). "Censored and shifted gamma distribution based EMOS model for probabilistic quantitative precipitation forecasting". *Environmetrics*, 27 (5), 280–292.

Baran, S., Szokol, P., and Szabó, M. (2021b). "Truncated generalized extreme value distribution-based ensemble model output statistics model for calibration of wind speed ensemble forecasts". *Environmetrics*, 32, e2678.

Bauer, P., Thorpe, A., and Brunet, G. (2015). "The quiet revolution of numerical weather prediction". *Nature*, 525 (7567), 47–55.

Becker, R. and Behrens, K. (2012). "Quality assessment of heterogeneous surface radiation network data". *Advances in Science and Research*, 8 (1), 93–97.

Benjamini, Y. and Hochberg, Y. (1995). "Controlling the false discovery rate: A practical and powerful approach to multiple testing". *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 57 (1), 289–300.

Bentzien, S. and Friederichs, P. (2012). "Generating and calibrating probabilistic quantitative precipitation forecasts from the high-resolution NWP Model COSMO-DE". *Weather and Forecasting*, 27 (4), 988–1002.

Berkowitz, J. (2001). "Testing density forecasts, with applications to risk management". *Journal of Business and Economic Statistics*, 19 (4), 465–474.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Tian, Q. (2022). *Pangu-Weather: A 3D high-resolution model for fast and accurate global weather forecast*. Preprint, available at `https://arxiv.org/abs/2211.02556`.

Bishop, C. M. (1994). *Mixture Density Networks*. Technical report, available at `https://publications.aston.ac.uk/id/eprint/373/1/NCRG_94_004.pdf`.

Boessenkool, B. (2021). *rdwd: Select and download climate data from 'DWD' (German Weather Service)*. R package version 1.5.0. `https://cran.r-project.org/package=rdwd`.

Bojer, C. S. and Meldgaard, J. P. (2021). "Kaggle forecasting competitions: An overlooked learning opportunity". *International Journal of Forecasting*, 37 (2), 587–603.

Bollmeyer, C., Keller, J. D., Ohlwein, C., Wahl, S., Crewell, S., Friederichs, P., Hense, A., Keune, J., Kneifel, S., Pscheidt, I., Redl, S., and Steinke, S. (2015). "Towards a high-resolution regional reanalysis for the european CORDEX domain". *Quarterly Journal of the Royal Meteorological Society*, 141 (686), 1–15.

Breiman, L. (1984). *Classification and Regression Trees*. Wadsworth International Group.

Breiman, L. (1996). "Bagging predictors". *Machine Learning*, 24 (2), 123–140.

Breiman, L. (2001). "Random forests". *Machine Learning*, 45 (1), 5–32.

Bremnes, J. B. (2020). "Ensemble postprocessing using quantile function regression based on neural networks and Bernstein polynomials". *Monthly Weather Review*, 148 (1), 403–414.

Brier, G. W. (1950). "Verification of forecasts expressed in terms of probability". *Monthly Weather Review*, 78 (1), 1–3.

Brown, B. G., Gilleland, E., and Ebert, E. E. (2011). "Forecasts of spatial fields". In: *Forecast Verification*. Wiley, 95–117.

Buizza, R. (1997). "Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF ensemble prediction system". *Monthly Weather Review*, 125 (1), 99–119.

Busetti, F. (2017). "Quantile aggregation of density forecasts". *Oxford Bulletin of Economics and Statistics*, 79 (4), 495–512.

Bustamante, J. (2017). *Bernstein Operators and Their Properties*. Springer.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). "A limited memory algorithm for bound constrained optimization". *SIAM Journal on Scientific Computing*, 16 (5), 1190–1208.

Chapman, W. E., Delle Monache, L., Alessandrini, S., Subramanian, A. C., Ralph, F. M., Xie, S.-P., Lerch, S., and Hayatbini, N. (2022). "Probabilistic predictions from deterministic atmospheric river forecasts with deep learning". *Monthly Weather Review*, 150 (1), 215–234.

Chen, J., Janke, T., Steinke, F., and Lerch, S. (2022). *Generative machine learning methods for multivariate ensemble post-processing*. Preprint, available at `http://arxiv.org/abs/2211.01345`.

Clare, M. C., Jamil, O., and Morcrette, C. J. (2021). "Combining distribution-based neural networks to predict weather forecast probabilities". *Quarterly Journal of the Royal Meteorological Society*, 147 (741), 4337–4357.

Cramer, E. Y. et al. (2022). "Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States". *Proceedings of the National Academy of Sciences of the United States of America*, 119 (15), e2113561119.

D'Isanto, A. and Polsterer, K. L. (2018). "Photometric redshift estimation via deep learning – Generalized and pre-classification-less, image based, fully probabilistic redshifts". *Astronomy and Astrophysics*, 609, A111.

Dai, Y. and Hemri, S. (2021). "Spatially coherent postprocessing of cloud cover ensemble forecasts". *Monthly Weather Review*, 149 (12), 3923–3937.

De Leeuw, J., Hornik, K., and Mair, P. (2009). "Isotone optimization in R: Pool-Adjacent-Violators Algorithm (PAVA) and active set methods". *Journal of Statistical Software*, 32 (5), 1–24.

Deutscher Wetterdienst (2018). *Ersetzung von COSMO-DE /-EPS durch COSMO-D2 /-EPS*. Technical report, available at `https://www.dwd.de/DE/fachnutzer/forschung_lehre/numerische_wettervorhersage/nwv_aenderungen/_functions/DownloadBox_modellaenderungen/cosmo_d2/pdf_2018_2020/pdf_cosmo_d2_15_05_2018.html`.

Deutscher Wetterdienst (2021). *Replacement of COSMO-D2 / COSMO-D2 EPS with ICON-D2 / ICON-D2-EPS*. Technical report, available at `https://www.dwd.de/DE/fachnutzer/forschung_lehre/numerische_wettervorhersage/nwv_aenderungen/_functions/DownloadBox_modellaenderungen/icon_d2/pdf_2021/pdf_icon_system_10_02_2021.html`.

Diebold, F. X. and Mariano, R. S. (1995). "Comparing predictive accuracy". *Journal of Business and Economic Statistics*, 13 (3), 253–263.

Dietterich, T. G. (2000). "Ensemble methods in machine learning". In: *Multiple Classifier Systems*. Lecture Notes in Computer Science, Vol. 1857. Springer, Berlin, Heidelberg.

Dimitriadis, T., Gneiting, T., and Jordan, A. I. (2021). "Stable reliability diagrams for probabilistic classifiers". *Proceedings of the National Academy of Sciences*, 118 (8), e2016191118.

Earl, N., Dorling, S., Starks, M., and Finch, R. (2017). "Subsynoptic-scale features associated with extreme surface gusts in UK extratropical cyclone events". *Geophysical Research Letters*, 44 (8), 3932–3940.

Eisenstein, L., Pantillon, F., and Knippertz, P. (2020). "Dynamics of sting-jet storm Egon over continental Europe: Impact of surface properties and model resolution". *Quarterly Journal of the Royal Meteorological Society*, 146 (726), 186–210.

Eisenstein, L., Schulz, B., Qadir, G. A., Pinto, J. G., and Knippertz, P. (2022). "Identification of high-wind features within extratropical cyclones using a probabilistic random forest – Part 1: Method and case studies". *Weather and Climate Dynamics*, 3 (4), 1157–1182.

Fan, R. E., Chang, K. W., Hsieh, C. J., Wang, X. R., and Lin, C. J. (2008). "LIBLINEAR: A library for large linear classification". *Journal of Machine Learning Research*, 9, 1871–1874.

Felder, M., Sehnke, F., Ohnmeiß, K., Schröder, L., Junk, C., and Kaifel, A. (2018). "Probabilistic short term wind power forecasts using deep neural networks with discrete target classes". *Advances in Geosciences*, 45, 13–17.

Fort, S., Hu, H., and Lakshminarayanan, B. (2019). *Deep ensembles: A loss landscape perspective*. Preprint, available at `https://doi.org/10.48550/arXiv.1912.02757`.

Fraley, C., Raftery, A. E., and Gneiting, T. (2010). "Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging". *Monthly Weather Review*, 138 (1), 190–202.

Fraunhofer Institute for Solar Energy Systems (2020). *Recent Facts about Photovoltaics in Germany.* Technical report, available at `https://www.ise.fraunhofer.de/en/publications/studies/recent-facts-about-pv-in-germany.html`.

Freund, Y. and Schapire, R. E. (1996). "Experiments with a new boosting algorithm". *Proceedings of the 13th International Conference on Machine Learning.* Citeseer, 148–156.

Gal, Y. and Ghahramani, Z. (2016). "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning". *Proceedings of the 33rd International Conference on Machine Learning.* PMLR 48, 1050–1059.

Gasthaus, J., Benidis, K., Wang, Y., Rangapuram, S. S., Salinas, D., Flunkert, V., and Januschowski, T. (2019). "Probabilistic forecasting with spline quantile function RNNs". *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics.* PMLR 89, 1901–1910.

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 Climatology, Meteorology and Oceanography*, 5 (1), 87–100.

Genest, C. (1992). "Vincentization revisited". *The Annals of Statistics*, 20 (2), 1137–1142.

Ghazvinian, M., Zhang, Y., Seo, D.-J., He, M., and Fernando, N. (2021). "A novel hybrid artificial neural network – Parametric scheme for postprocessing medium-range precipitation forecasts". *Advances in Water Resources*, 151, 103907.

Giorgi, F., Jones, C., and Asrar, G. R. (2009). "Addressing climate information needs at the regional level: The CORDEX framework". *World Meteorological Organization (WMO) Bulletin*, 58 (3), 175.

Gneiting, T. (2011). "Making and evaluating point forecasts". *Journal of the American Statistical Association*, 106 (494), 746–762.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). "Probabilistic forecasts, calibration and sharpness". *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 69 (2), 243–268.

Gneiting, T. and Katzfuss, M. (2014). "Probabilistic forecasting". *Annual Review of Statistics and Its Application*, 1 (1), 125–151.

Gneiting, T., Larson, K., Westrick, K., Genton, M. G., and Aldrich, E. (2006). "Calibrated probabilistic forecasting at the Stateline wind energy center: The regime-switching space-time method". *Journal of the American Statistical Association*, 101 (475), 968–979.

Gneiting, T., Lerch, S., and Schulz, B. (2023). "Probabilistic solar forecasting: Benchmarks, post-processing, verification". *Solar Energy*, 252, 72–80.

Gneiting, T. and Raftery, A. E. (2007). "Strictly proper scoring rules, prediction, and estimation". *Journal of the American Statistical Association*, 102 (477), 359–378.

Gneiting, T., Raftery, A. E., Westveld, A. H., and Goldman, T. (2005). "Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation". *Monthly Weather Review*, 133 (5), 1098–1118.

Gneiting, T. and Ranjan, R. (2011). "Comparing density forecasts using threshold and quantile-weighted scoring rules". *Journal of Business and Economic Statistics*, 29 (3), 411–422.

Gneiting, T. and Ranjan, R. (2013). "Combining predictive distributions". *Electronic Journal of Statistics*, 7 (1), 1747–1782.

González Ordiano, J. Á., Gröll, L., Mikut, R., and Hagenmeyer, V. (2020). "Probabilistic energy forecasting using the nearest neighbors quantile filter and quantile regression". *International Journal of Forecasting*, 36 (2), 310–323.

Good, I. J. (1952). "Rational decisions". *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 14 (1), 107–114.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

Gottwalt, S., Gärttner, J., Schmeck, H., and Weinhardt, C. (2017). "Modeling and valuation of residential demand flexibility for renewable energy integration". *IEEE Transactions on Smart Grid*, 8 (6), 2565–2574.

Gray, S. L., Martínez-Alvarado, O., Ackerley, D., and Suri, D. (2021). "Development of a prototype real-time sting-jet precursor tool for forecasters". *Weather*, 76 (11), 369–373.

Greenwell, B. M. (2017). "pdp: An R package for constructing partial dependence plots". *R Journal*, 9 (1), 421–436.

Gregory, P. A., Camp, J., Bigelow, K., and Brown, A. (2019). "Sub-seasonal predictability of the 2017–2018 Southern Hemisphere tropical cyclone season". *Atmospheric Science Letters*, 20 (4), e886.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). "On calibration of modern neural networks". *Proceedings of the 34th International Conference on Machine Learning*. PMLR 70, 1321–1330.

Guttorp, P. and Gneiting, T. (2006). "Studies in the history of probability and statistics XLIX on the Matérn correlation family". *Biometrika*, 93 (4), 989–995.

Hamill, T. M., Hagedorn, R., and Whitaker, J. S. (2008). "Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation". *Monthly Weather Review*, 136 (7), 2620–2632.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.

Haupt, S. E., Chapman, W., Adams, S. V., Kirkwood, C., Hosking, J. S., Robinson, N. H., Lerch, S., and Subramanian, A. C. (2021). "Towards implementing artificial intelligence post-processing in weather and climate: Proposed actions from the Oxford 2019 workshop". *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379 (2194), 20200091.

Haupt, S. E., Garcia Casado, M., Davidson, M., Dobschinski, J., Du, P., Lange, M., Miller, T., Mohrlen, C., Motley, A., Pestana, R., and Zack, J. (2019). "The use of probabilistic forecasts: Applying them in theory and practice". *IEEE Power and Energy Magazine*, 17 (6), 46–57.

Haupt, S. E., McCandless, T. C., Dettling, S., Alessandrini, S., Lee, J. A., Linden, S., Petzke, W., Brummet, T., Nguyen, N., Kosović, B., Wiener, G., Hussain, T., and Al-Rasheedi, M. (2020). "Combining artificial intelligence with physics-based methods for probabilistic renewable energy forecasting". *Energies*, 13 (8), 1979.

Hemri, S., Haiden, T., and Pappenberger, F. (2016). "Discrete postprocessing of total cloud cover ensemble forecasts". *Monthly Weather Review*, 144 (7), 2565–2577.

Hemri, S., Scheuerer, M., Pappenberger, F., Bogner, K., and Haiden, T. (2014). "Trends in the predictive performance of raw ensemble weather forecasts". *Geophysical Research Letters*, 41 (24), 9197–9205.

Henzi, A., Kleger, G.-R., and Ziegel, J. F. (2021a). "Distributional (single) index models". *Journal of the American Statistical Association*.

Henzi, A., Ziegel, J. F., and Gneiting, T. (2019). *isodistrreg: Isotonic Distributional Regression (IDR)*. R package version 0.1.0.9000. `https://github.com/AlexanderHenzi/isodistrreg`.

Henzi, A., Ziegel, J. F., and Gneiting, T. (2021b). "Isotonic distributional regression". *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 83 (5), 963–993.

Hersbach, H. et al. (2020). "The ERA5 global reanalysis". *Quarterly Journal of the Royal Meteorological Society*, 146 (730), 1999–2049.

Hess, R. (2020). "Statistical postprocessing of ensemble forecasts for severe weather at Deutscher Wetterdienst". *Nonlinear Processes in Geophysics*, 27 (4), 473–487.

Hewson, T. D. and Neu, U. (2015). "Cyclones, windstorms and the IMILAST project". *Tellus A: Dynamic Meteorology and Oceanography*, 6 (1), 27128.

Hill, A. J., Herman, G. R., and Schumacher, R. S. (2020). "Forecasting severe weather with random forests". *Monthly Weather Review*, 148 (5), 2135–2161.

Hong, T. and Fan, S. (2016). "Probabilistic electric load forecasting: A tutorial review". *International Journal of Forecasting*, 32 (3), 914–938.

Hong, T., Pinson, P., Wang, Y., Weron, R., Yang, D., and Zareipour, H. (2020). "Energy forecasting: A review and outlook". *IEEE Open Access Journal of Power and Energy*, 7, 376–388.

Januschowski, T., Wang, Y., Torkkola, K., Erkkilä, T., Hasson, H., and Gasthaus, J. (2022). "Forecasting with trees". *International Journal of Forecasting*, 38 (4), 1473–1481.

Jávorné Radnóczi, K., Várkonyi, A., and Szépszó, G. (2020). *On the way towards the AROME nowcasting system in Hungary*. ALADIN-HIRLAM Newsletter, 14, 65–69.

Jordan, A. (2016). *Facets of Forecast Evaluation*. PhD thesis, Karlsruhe Institute of Technology.

Jordan, A., Krüger, F., and Lerch, S. (2019). "Evaluating probabilistic forecasts with scoringRules". *Journal of Statistical Software*, 90 (12), 1–37.

Jospin, L. V., Laga, H., Boussaid, F., Buntine, W., and Bennamoun, M. (2022). "Hands-on Bayesian neural networks – A tutorial for deep learning users". *IEEE Computational Intelligence Magazine*, 17 (2), 29–48.

Keisler, R. (2022). *Forecasting global weather with graph neural networks*. Preprint, available at `https://arxiv.org/abs/2202.07575`.

Keller, R., Rajczak, J., Bhend, J., Spirig, C., Hemri, S., Liniger, M. A., and Wernli, H. (2021). "Seamless multi-model postprocessing for air temperature forecasts in complex topography". *Weather and Forecasting*, 36 (3), 1031–1042.

Kim, T., Fakoor, R., Mueller, J., Smola, A. J., and Tibshirani, R. J. (2021). *Deep quantile aggregation*. Preprint, available at `http://arxiv.org/abs/2103.00083`.

Kingma, D. and Ba, J. (2014). *Adam: A method for stochastic optimization*. Preprint, available at `https://arxiv.org/abs/1412.6980v8`.

Kirkwood, C., Economou, T., Odbert, H., and Pugeault, N. (2021). "A framework for probabilistic weather forecast post-processing across models and lead times using machine learning". *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379 (2194).

Klawa, M. and Ulbrich, U. (2003). "A model for the estimation of storm losses and the identification of severe winter storms in Germany". *Natural Hazards and Earth System Science*, 3 (6), 725–732.

Kleczek, M. A., Steeneveld, G.-J., and Holtslag, A. A. M. (2014). "Evaluation of the weather research and forecasting mesoscale model for GABLS3: Impact of boundary-layer schemes, boundary conditions and spin-up". *Boundary-Layer Meteorology*, 152 (2), 213–243.

Knapp, K. R., Diamond, H. J., Kossin, J. P., Kruk, M. C., and Schreck, C. J. (2018). *International best track archive for climate stewardship (IBTrACS) project, version 4*. NOAA/National Centers for Environmental Information, accessed 14/04/2020.

Knapp, K. R., Kruk, M. C., Levinson, D. H., Diamond, H. J., and Neumann, C. J. (2010). "The international best track archive for climate stewardship (IBTrACS)". *Bulletin of the American Meteorological Society*, 91 (3), 363–376.

Koliander, G., El-Laham, Y., Djuric, P. M., and Hlawatsch, F. (2022). "Fusion of probability density functions". *Proceedings of the IEEE*, 110 (4), 404–453.

Krüger, F., Lerch, S., Thorarinsdottir, T., and Gneiting, T. (2021). "Predictive inference based on Markov chain Monte Carlo output". *International Statistical Review*, 89 (2), 274–301.

Kull, M., Filho, T. M. S., and Flach, P. (2017). "Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration". *Electronic Journal of Statistics*, 11 (2), 5052–5080.

Lackmann, G. (2011). "Chapter 6: Fronts". In: *Midlatitude Synoptic Meteorology: Dynamics, Analysis and Forecasting*. Boston, Mass.: American Meteorological Society, 131–166.

Lagerquist, R., McGovern, A., and Smith, T. (2017). "Machine learning for real-time prediction of damaging straight-line convective wind". *Weather and Forecasting*, 32 (6), 2175–2193.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). "Simple and scalable predictive uncertainty estimation using deep ensembles". *Advances in Neural Information Processing Systems*. NIPS 2017, 6405–6414.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Pritzel, A., Ravuri, S., Ewalds, T., Alet, F., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Stott, J., Vinyals, O., Mohamed, S., and Battaglia, P. (2022). *GraphCast: Learning skillful medium-range global weather forecasting*. Preprint, available at `https://arxiv.org/abs/2212.12794`.

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 (1), 23–34.

Le Gal La Salle, J., Badosa, J., David, M., Pinson, P., and Lauret, P. (2020). "Added-value of ensemble prediction system on the quality of solar irradiance probabilistic forecasts". *Renewable Energy*, 162, 1321–1339.

Lee, C. Y., Camargo, S. J., Vitart, F., Sobel, A. H., Camp, J., Wang, S., Tippett, M. K., and Yang, Q. (2020). "Subseasonal predictions of tropical cyclone occurrence and ace in the S2S dataset". *Weather and Forecasting*, 35 (3), 921–938.

Lee, C. Y., Camargo, S. J., Vitart, F., Sobel, A. H., and Tippett, M. K. (2018). "Subseasonal tropical cyclone genesis prediction and MJO in the S2S dataset". *Weather and Forecasting*, 33 (4), 967–988.

Lee, S., Purushwalkam, S., Cogswell, M., Crandall, D., and Batra, D. (2015). *Why M heads are better than one: Training a diverse ensemble of deep networks*. Preprint, available at `https://arxiv.org/abs/1511.06314v1`.

Lerch, S. and Baran, S. (2017). "Similarity-based semilocal estimation of post-processing models". *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 66 (1), 29–51.

Lerch, S., Baran, S., Möller, A., Groß, J., Schefzik, R., Hemri, S., and Graeter, M. (2020). "Simulation-based comparison of multivariate ensemble post-processing methods". *Nonlinear Processes in Geophysics*, 27 (2), 349–371.

Lerch, S. and Thorarinsdottir, T. L. (2013). "Comparison of non-homogeneous regression models for probabilistic wind speed forecasting". *Tellus A: Dynamic Meteorology and Oceanography*, 65 (1), 21206.

Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F., and Gneiting, T. (2017). "Forecaster's dilemma: Extreme events and forecast evaluation". *Statistical Science*, 32 (1), 106–127.

Leroy, A. and Wheeler, M. C. (2008). "Statistical prediction of weekly tropical cyclone activity in the Southern Hemisphere". *Monthly Weather Review*, 136 (10), 3637–3654.

Li, R., Reich, B. J., and Bondell, H. D. (2021). "Deep distribution regression". *Computational Statistics and Data Analysis*, 159, 107203.

Lichtendahl, K. C., Grushka-Cockayne, Y., and Winkler, R. L. (2013). "Is it better to average probabilities or quantiles?" *Management Science*, 59 (7), 1594–1611.

Lorenz, E. N. (1963). "Deterministic nonperiodic flow". *Journal of the Atmospheric Sciences*, 20 (2), 130–141.

Lundstrom, L. (2016). *camsRad: Client for CAMS radiation service*. R package version 0.3.0. `https://cran.r-project.org/package=camsRad`.

Mahrt, L. (2017). "The near-surface evening transition". *Quarterly Journal of the Royal Meteorological Society*, 143 (708), 2940–2948.

Maier-Gerber, M., Fink, A. H., Riemer, M., Schoemer, E., Fischer, C., and Schulz, B. (2021). "Statistical-dynamical forecasting of sub-seasonal North Atlantic tropical cyclone occurrence". *Weather and Forecasting*, 36 (6), 2127–2142.

Matérn, B. (1986). *Spatial Variation*. 2nd. Lecture Notes in Statistics, Vol. 36. Berlin: Springer-Verlag.

Matheron, G. (1963). "Principles of geostatistics". *Economic Geology*, 58 (8), 1246–1266.

Matheson, J. E. and Winkler, R. L. (1976). "Scoring rules for continuous probability distributions". *Management Science*, 22 (10), 1087–1096.

McGovern, A., Lagerquist, R., Gagne, D. J., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., and Smith, T. (2019). "Making the black box more transparent: Understanding the physical implications of machine learning". *Bulletin of the American Meteorological Society*, 100 (11), 2175–2199.

Meinshausen, N. (2006). "Quantile regression forests". *Journal of Machine Learning Research*, 7, 983–999.

Merz, B., Kuhlicke, C., Kunz, M., Pittore, M., Babeyko, A., Bresch, D. N., Domeisen, D. I., Feser, F., Koszalka, I., Kreibich, H., Pantillon, F., Parolai, S., Pinto, J. G., Punge, H. J., Rivalta, E., Schröter, K., Strehlow, K., Weisse, R., and Wurpts, A. (2020). "Impact forecasting to support emergency management of natural hazards". *Reviews of Geophysics*, 58 (4), e2020RG000704.

Messner, J. W., Mayr, G. J., and Zeileis, A. (2016). "Heteroscedastic censored and truncated regression with crch". *R Journal*, 8 (1), 173–181.

Messner, J. W., Mayr, G. J., and Zeileis, A. (2017). "Nonhomogeneous boosting for predictor selection in ensemble postprocessing". *Monthly Weather Review*, 145 (1), 137–147.

Mohamed, S. and Lakshminarayanan, B. (2016). *Learning in implicit generative models*. Preprint, available at `https://arxiv.org/abs/1610.03483`.

Molnar, C. (2018). *Interpretable Machine Learning*. Available at `https://christophm.github.io/interpretable-ml-book/`.

Molteni, F., Buizza, R., Palmer, T. N., and Petroliagis, T. (1996). "The ECMWF ensemble prediction system: Methodology and validation". *Quarterly Journal of the Royal Meteorological Society*, 122 (529), 73–119.

Murphy, A. H. and Winkler, R. L. (1987). "A general framework for forecast verification". *Monthly Weather Review*, 115 (7), 1330–1338.

Neal, R. M. (2012). *Bayesian Learning for Neural Networks*. Lecture Notes in Statistics, Vol. 118. Springer New York.

Osband, I., Wen, Z., Asghari, S. M., Dwaracherla, V., Hao, B., Ibrahimi, M., Lawson, D., Lu, X., O'Donoghue, B., and van Roy, B. (2021). *Evaluating predictive distributions: Does Bayesian deep learning work?* Preprint, available at `https://doi.org/10.48550/arXiv.2110.04629`.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., and Snoek, J. (2019). "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift". *Advances in Neural Information Processing Systems*. NeurIPS 2019.

Pan, S. J. and Yang, Q. (2010). "A survey on transfer learning". *IEEE Transactions on Knowledge and Data Engineering*, 22 (10), 1345–1359.

Pantillon, F., Lerch, S., Knippertz, P., and Corsmeier, U. (2018). "Forecasting wind gusts in winter storms using a calibrated convection-permitting ensemble". *Quarterly Journal of the Royal Meteorological Society*, 144 (715), 1864–1881.

Parton, G., Dore, A., and Vaughan, G. (2010). "A climatology of mid-tropospheric mesoscale strong wind events as observed by the MST radar, Aberystwyth". *Meteorological Applications*, 17 (3), 340–354.

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., Kurth, T., Hall, D., Li, Z., Azizzadenesheli, K., Hassanzadeh, P., Kashinath, K., and Anandkumar, A. (2022). *FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators*. Preprint, available at `https://arxiv.org/abs/2202.11214`.

Peterson, R. A. (2021). "Finding optimal normalizing transformations via bestNormalize". *R Journal*, 13 (1), 310–329.

Petropoulos, F. et al. (2022). "Forecasting: Theory and practice". *International Journal of Forecasting*, 38 (3), 705–871.

Phipps, K., Lerch, S., Andersson, M., Mikut, R., Hagenmeyer, V., and Ludwig, N. (2022). "Evaluating ensemble post-processing for wind power forecasts". *Wind Energy*, 25 (8), 1379–1405.

Pinson, P. and Messner, J. W. (2018). "Application of postprocessing for renewable energy". In: *Statistical Postprocessing of Ensemble Forecasts*. Ed. by S. Vannitsem, D. S. Wilks, and J. W. Messner. Elsevier, 241–266.

Platt, J. (1999). "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods". *Advances in large margin classifiers*, 10 (3), 61–74.

Politis, D. N. and Romano, J. P. (1994). "The stationary bootstrap". *Journal of the American Statistical Association*, 89 (428), 1303–1313.

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. `https://www.R-project.org/`.

Rahaman, R. and Thiery, A. H. (2021). "Uncertainty quantification and deep ensembles". *Advances in Neural Information Processing Systems*. NeurIPS 2021, 20063–20075.

Ranjan, R. and Gneiting, T. (2010). "Combining probability forecasts". *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 72 (1), 71–91.

Rasmussen, C. E. and Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning*. The MIT Press.

Rasp, S. and Lerch, S. (2018). "Neural networks for postprocessing ensemble weather forecasts". *Monthly Weather Review*, 146 (11), 3885–3900.

Ratcliff, R. (1979). "Group reaction time distributions and an analysis of distribution statistics". *Psychological Bulletin*, 86 (3), 446–461.

Ren, Y., Zhang, L., and Suganthan, P. N. (2016). "Ensemble classification and regression – Recent developments, applications and future directions". *IEEE Computational Intelligence Magazine*, 11 (1), 41–53.

Robertson, A. W., Vitart, F., and Camargo, S. J. (2020). "Subseasonal to seasonal prediction of weather to climate with application to tropical cyclones". *Journal of Geophysical Research: Atmospheres*, 125 (6), e2018JD029375.

Rodwell, M. J., Richardson, D. S., Parsons, D. B., and Wernli, H. (2018). "Flow-dependent reliability: A path to more skillful ensemble forecasts". *Bulletin of the American Meteorological Society*, 99 (5), 1015–1026.

Sanders, F. (1963). "On subjective probability forecasting". *Journal of Applied Meteorology*, 2 (2), 191–201.

Satopää, V. A., Baron, J., Foster, D. P., Mellers, B. A., Tetlock, P. E., and Ungar, L. H. (2014). "Combining multiple probability predictions using a simple logit model". *International Journal of Forecasting*, 30 (2), 344–356.

Schefzik, R. (2017). "Ensemble calibration with preserved correlations: Unifying and comparing ensemble copula coupling and member-by-member postprocessing". *Quarterly Journal of the Royal Meteorological Society*, 143 (703), 999–1008.

Schefzik, R., Thorarinsdottir, T. L., and Gneiting, T. (2013). "Uncertainty quantification in complex simulation models using ensemble copula coupling". *Statistical Science*, 28 (4), 616–640.

Scheuerer, M. (2014). "Probabilistic quantitative precipitation forecasting using ensemble model output statistics". *Quarterly Journal of the Royal Meteorological Society*, 140 (680), 1086–1096.

Scheuerer, M. and Möller, D. (2015). "Probabilistic wind speed forecasting on a grid based on ensemble model output statistics". *Annals of Applied Statistics*, 9 (3), 1328–1349.

Scheuerer, M., Switanek, M. B., Worsnop, R. P., and Hamill, T. M. (2020). "Using artificial neural networks for generating probabilistic subseasonal precipitation forecasts over California". *Monthly Weather Review*, 148 (8), 3489–3506.

Schraff, C., Reich, H., Rhodin, A., Schomburg, A., Stephan, K., Periáñez, A., and Potthast, R. (2016). "Kilometre-scale ensemble data assimilation for the COSMO model (KENDA)". *Quarterly Journal of the Royal Meteorological Society*, 142 (696), 1453–1472.

Schultz, M. G., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L. H., Mozaffari, A., and Stadtler, S. (2021). "Can deep learning beat numerical weather prediction?" *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379 (2194).

Schulz, B., El Ayari, M., Lerch, S., and Baran, S. (2021). "Post-processing numerical weather prediction ensembles for probabilistic solar irradiance forecasting". *Solar Energy*, 220, 1016–1031.

Schulz, B. and Lerch, S. (2022a). "Aggregating distribution forecasts from deep ensembles". *Journal of Machine Learning Research*, under revision. Preprint available at `https://arxiv.org/abs/2204.02291v1`.

Schulz, B. and Lerch, S. (2022b). "Machine learning methods for postprocessing ensemble forecasts of wind gusts: A systematic comparison". *Monthly Weather Review*, 150 (1), 235–257.

Skok, G. and Hladnik, V. (2018). "Verification of gridded wind forecasts in complex Alpine terrain: A new wind verification methodology based on the neighborhood approach". *Monthly Weather Review*, 146 (1), 63–75.

Slade, S. A. and Maloney, E. D. (2013). "An intraseasonal prediction model of Atlantic and east Pacific tropical cyclone genesis". *Monthly Weather Review*, 141 (6), 1925–1942.

Sloughter, J. M. L., Raftery, A. E., Gneiting, T., and Fraley, C. (2007). "Probabilistic quantitative precipitation forecasting using Bayesian model averaging". *Monthly Weather Review*, 135 (9), 3209–3220.

Sperati, S., Alessandrini, S., and Delle Monache, L. (2016). "An application of the ECMWF Ensemble Prediction System for short-term solar power forecasting". *Solar Energy*, 133, 437–450.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). "Dropout: A simple way to prevent neural networks from overfitting". *Journal of Machine Learning Research*, 15 (1), 1929–1958.

Stone, M. (1961). "The opinion pool". *The Annals of Mathematical Statistics*, 32 (4), 1339–1342.

Taillardat, M. and Mestre, O. (2020). "From research to applications – Examples of operational ensemble post-processing in France using machine learning". *Nonlinear Processes in Geophysics*, 27 (2), 329–347.

Taillardat, M., Mestre, O., Zamo, M., and Naveau, P. (2016). "Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics". *Monthly Weather Review*, 144 (6), 2375–2393.

Taylor, J. W. and Taylor, K. S. (2023). "Combining probabilistic forecasts of COVID-19 mortality in the United States". *European Journal of Operational Research*, 304 (1), 25–41.

Thomas, E. A. and Ross, B. H. (1980). "On appropriate procedures for combining probability distributions within the same family". *Journal of Mathematical Psychology*, 21 (2), 136–152.

Thorarinsdottir, T. L. and Gneiting, T. (2010). "Probabilistic forecasts of wind speed: Ensemble model output statistics by using heteroscedastic censored regression". *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 173 (2), 371–388.

Toth, Z. and Kalnay, E. (1993). "Ensemble forecasting at NMC: The generation of perturbations". *Bulletin of the American Meteorological Society*, 74 (12), 2317–2330.

Van der Meer, D. W., Widén, J., and Munkhammar, J. (2018). "Review on probabilistic forecasting of photovoltaic power production and electricity consumption". *Renewable and Sustainable Energy Reviews*, 81, 1484–1512.

Van Schaeybroeck, B. and Vannitsem, S. (2015). "Ensemble post-processing using member-by-member approaches: Theoretical aspects". *Quarterly Journal of the Royal Meteorological Society*, 141 (688), 807–818.

Vannitsem, S., Wilks, D. S., and Messner, J. W. (2018). *Statistical Postprocessing of Ensemble Forecasts*. Elsevier.

Vannitsem, S. et al. (2021). "Statistical postprocessing for weather forecasts: Review, challenges, and avenues in a big data world". *Bulletin of the American Meteorological Society*, 102 (3), E681–E699.

Veldkamp, S., Whan, K., Dirksen, S., and Schmeits, M. (2021). "Statistical postprocessing of wind speed forecasts using convolutional neural networks". *Monthly Weather Review*, 149 (4), 1141–1152.

Vincent, S. B. (1912). "The functions of the Vibrissae in the behavior of the white rat". *Animal Behavior Monographs*, 1 (5).

Vitart, F. et al. (2017). "The subseasonal to seasonal (S2S) prediction project database". *Bulletin of the American Meteorological Society*, 98 (1), 163–173.

Vitart, F., Leroy, A., and Wheeler, M. C. (2010). "A comparison of dynamical and statistical predictions of weekly tropical cyclone activity in the Southern Hemisphere". *Monthly Weather Review*, 138 (9), 3671–3682.

Vitart, F., Robertson, A. W., and Anderson, D. L. T. (2012). "Subseasonal to seasonal prediction project: Bridging the gap between weather and climate". *World Meteorological Organization (WMO) Bulletin*, 61 (2), 23–28.

Vogel, P., Knippertz, P., Fink, A. H., Schlueter, A., and Gneiting, T. (2018). "Skill of global raw and postprocessed ensemble predictions of rainfall over northern tropical Africa". *Weather and Forecasting*, 33 (2), 369–388.

Walz, E., Maranan, M., van der Linden, R., Fink, A. H., and Knippertz, P. (2021). "An IMERG-based optimal Extended Probabilistic Climatology (EPC) as a benchmark ensemble forecast for precipitation in the tropics and subtropics". *Weather and Forecasting*, 36 (4), 1561–1573.

Wang, J. and Ghosh, S. K. (2012). "Shape restricted nonparametric regression with Bernstein polynomials". *Computational Statistics and Data Analysis*, 56 (9), 2729–2741.

Warner, T. T. (2010). *Numerical Weather and Climate Prediction*. Cambridge.

Weiss, K., Khoshgoftaar, T. M., and Wang, D. (2016). "A survey of transfer learning". *Journal of Big Data*, 3 (1), 9.

Wen, Y., Tran, D., and Ba, J. (2020). *BatchEnsemble: An alternative approach to efficient ensemble and lifelong learning*. Preprint, available at `http://arxiv.org/abs/2002.06715`.

Wilks, D. S. (2016). "'The stippling shows statistically significant grid points': How research results are routinely overstated and overinterpreted, and what to do about it". *Bulletin of the American Meteorological Society*, 97 (12), 2263–2273.

Wilks, D. S. (2018). "Univariate ensemble postprocessing". In: *Statistical Postprocessing of Ensemble Forecasts*. Ed. by S. Vannitsem, D. S. Wilks, and J. Messner. Elsevier, 49–89.

Wohland, J., Reyers, M., Märker, C., and Witthaut, D. (2018). "Natural wind variability triggered drop in German redispatch volume and costs from 2015 to 2016". *PLOS ONE*, 13 (1), 1–21.

Wolffram, D. (2021). *Building and Evaluating Forecast Ensembles for COVID-19 Deaths*. M.Sc. thesis, Karlsruhe Institute of Technology.

Wright, M. N. and Ziegler, A. (2017). "Ranger: A fast implementation of random forests for high dimensional data in C++ and R". *Journal of Statistical Software*, 77 (1), 1–17.

Wu, X. and Gales, M. (2021). *Should ensemble members be calibrated?* Preprint, available at `http://arxiv.org/abs/2101.05397`.

Yagli, G. M., Yang, D., and Srinivasan, D. (2020). "Ensemble solar forecasting using data-driven models with probabilistic post-processing through GAMLSS". *Solar Energy*, 208, 612–622.

Yang, D. (2019). "A guideline to solar forecasting research practice: Reproducible, operational, probabilistic or physically-based, ensemble, and skill (ROPES)". *Journal of Renewable and Sustainable Energy*, 11 (2), 22701.

Yang, D. (2020a). "Choice of clear-sky model in solar forecasting". *Journal of Renewable and Sustainable Energy*, 12 (2), 26101.

Yang, D. (2020b). "Ensemble model output statistics as a probabilistic site-adaptation tool for solar irradiance: A revisit". *Journal of Renewable and Sustainable Energy*, 12 (3), 36101.

Yang, D., Wang, W., and Hong, T. (2022). "A historical weather forecast dataset from the European Centre for Medium-Range Weather Forecasts (ECMWF) for energy forecasting". *Solar Energy*, 232, 263–274.

Yang, D. et al. (2020). "Verification of deterministic solar forecasts". *Solar Energy*, 210, 20–37.

Zadrozny, B. and Elkan, C. (2002). "Transforming classifier scores into accurate multiclass probability estimates". *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, 694–699.

Zängl, G., Reinert, D., Rípodas, P., and Baldauf, M. (2015). "The ICON (ICOsahedral Non-hydrostatic) modelling framework of DWD and MPI-M: Description of the non-hydrostatic dynamical core". *Quarterly Journal of the Royal Meteorological Society*, 141 (687), 563–579.

Zelikman, E., Zhou, S., Irvin, J., Raterink, C., Sheng, H., Avati, A., Kelly, J., Rajagopal, R., Ng, A. Y., and Gagne, D. (2020). *Short-term solar irradiance forecasting using calibrated probabilistic models*. Preprint, available at `https://arxiv.org/abs/2010.04715`.

Zhou, Z.-H., Wu, J., and Tang, W. (2002). "Ensembling neural networks: Many could be better than all". *Artificial Intelligence*, 137 (1-2), 239–263.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. (2021). "A comprehensive survey on transfer learning". *Proceedings of the IEEE*, 109 (1), 43–76.