Post-processing of Ensemble Forecasts of Various Weather and Hydrological Quantities Using Statistical and Machine Learning Methods

Thesis for the Degree of Doctor of Philosophy (PhD)

by Mehrez El Ayari

Supervisor:
Prof. Dr. Sándor Baran

UNIVERSITY OF DEBRECEN
Doctoral Council of Natural Sciences and Information Technology
Doctoral School of Informatics
Debrecen, 2022

Hereby I declare that I prepared this thesis within the Doctoral Council of Natural Sciences and Information Technology, Doctoral School of Informatics, University of Debrecen in order to obtain a PhD Degree in Informatics at Debrecen University.

The results published in the thesis are not reported in any other PhD theses.

Debrecen, 202. . . . . .

signature of the candidate

Hereby I confirm that Mehrez El Ayari candidate conducted his studies with my supervision within the Theoretical Foundation and Applications of Information Technology and Stochastic Systems Doctoral Program of the Doctoral School of Informatics between 2017 and 2022. The independent studies and research work of the candidate significantly contributed to the results published in the thesis.
I also declare that the results published in the thesis are not reported in any other PhD theses.
I support the acceptance of the thesis.

Debrecen, 202. . . . . .

signature of the supervisor

Post-processing of Ensemble Forecasts of Various Weather and Hydrological Quantities Using Statistical and Machine Learning Methods

Dissertation submitted in partial fulfilment of the requirements
for the doctoral (PhD) degree in Informatics

Written by Mehrez El Ayari, certified Hydrologist and Environmental scientist

Prepared in the framework of the informatics doctoral school of the
University of Debrecen

(Theoretical Foundation and Applications of Information Technology and
Stochastic Systems programme)

Dissertation advisor: Prof. Dr. Sándor Baran

The official opponents of the dissertation:

Dr. ....
Dr. ....
Dr. ....

The evaluation committee:

chairperson: Dr. ....
members: Dr. ....
Dr. ....
Dr. ....
Dr. ....

The date of the dissertation defence: ..... 20...

List of abbreviations

AROME-EPS	Applications of Research to Operations at Mesoscale- Ensemble Prediction Systems
BMA	Bayesian Model Averaging
BS	Brier Score
BfG	German Federal Institute of Hydrology
CRPS	Continuous ranked probability score
CRPSS	Continuous ranked probability skill score
CTRL	Control Forecasts
COSMO LEPS	The Consortium for Small-scale Modelling Limited area Ensemble Prediction System
CART	Classification and regression trees
DM	Diebold-Mariano
DHI	Diffuse horizontal radiation
DNI	Direct normal irradiance
DWD	Deutsche Wetterdienst
DTW	Dynamic Time Wraping
ECMWF	European Center for Medium-Range Weather Forecasts
EDA	Ensemble of Data Assimilations
EMOS	Ensemble Model Output Statistics
EPS	Ensemble Prediction Systems
ENS	the ECMWF medium-range/monthly ensemble
EM	Expectation–Maximization

GBM . . . . . Gradient Boosting Machines

GBMS . . . . . Gradient Boosting Machines Seasonal

GBMS-P . . . . Gradient Boosting Machines Seasonal - Precipitation

GHI . . . . . Global horizontal irradiance

HMS . . . . . Hungarian Meteorological Service

HMA . . . . . Hydrograph Matching Algorithm

HBV . . . . . Hydrologiska Byråns Vattenbalansavdelning

HRES . . . . . the ECMWF (single) high-resolution forecast

ICON-EPS . . . ICOsahedralNonhydrostatic- Ensemble Prediction System

LogS . . . . . Logarithmic Score

LogSS . . . . . Logarithmic skill score

MLR . . . . . Multiclass Logistic Regression

MLRS . . . . . Multiclass Logistic Regression Seasonal

MLRS-P . . . . Multiclass Logistic Regression Seasonal - Precipitation

MLP . . . . . Multilayer perceptron

MLPS . . . . . Multilayer perceptron Seasonal

MLPS-P . . . . Multilayer perceptron Seasonal - Precipitation

ML . . . . . Maximum likelihood

MET . . . . . Monthly expanding training

NWP . . . . . Numerical Weather Prediction

NR . . . . . Non-homogeneous Regression

NCEP GEFS . The National Centers for Environmental Prediction , Global Ensemble Forecast System.

PMF . . . . . Probability Mass Function

POLR . . . . . Proportional Odds Logistic Regression

POLRS . . . . . Proportional Odds Logistic Regression Seasonal

POLRS-P . . . . Proportional Odds Logistic Regression Seasonal - Precipitation

PV . . . . . Photovoltaic

PIT . . . . . probability integral transform

QRF . . . . . Quantile regression forest
RTP . . . . . Rolling Training Periods
SD . . . . . Series Distance
SKEB . . . . . Stochastic kinetic energy backscatter
SPPT . . . . . Stochastically perturbed parametrization tendency
SVs . . . . . Singular vectors
SYNOP . . . . . Surface synoptic observations
TCC . . . . . Total Cloud Cover

Contents

List of Figures

List of Tables

Introduction

Since dawn of time, humans are bound to weather conditions. To meet the needs of inhabitants, most of great ancient cities were geographically placed in intertropical areas, because the agricultural production is heavily depending on the climate. Therefore, extreme weather events can lead to humanitarian crises. In modern society, such events have both social and economic impacts.

Nowadays accurate weather forecasts became indispensable for many areas such as renewable energies, management of freight transport, and natural disaster control. With the advent of computers and supercomputers, one now is able to produce Numerical Weather Prediction (NWP) providing forecast variables (ensemble forecasts) which we can use as predictors for weather forecasting. Those ensemble forecasts issued differently either in the numerical model or its initial condition. Despite that those forecasts are subject to errors and not reliable in terms of representation of the weather quantity at hand. The necessissity to reduce those errors or increase the sharpness of those forecasts has created the need to develop a new field: "ensemble post-processing".

Post-processing adds value to the NWP model output by improving the sharpness and performance. Post-processing takes the form of any systematic altering of the model output before it is publicly made available. It is basically a statistical model that relates the observed variables of interest with the corresponding model output. Given the ensemble forecasts and their corresponding observations for a given lead time and training period, one can derive distribution based model. According to the dataset, the training period can be of the form of rolling window or of the past or a sliding window around a specific time. After that we have a post-processor that corrects the biases and dispersion errors, preserves the predictive skill and the statistical dependency structure of space and time of the ensemble forecasts (Schaake et al., 2007; Gneiting et al., 2007; Gneiting and Katzfuss, 2014; Yuan et al., 2015). Additional benefit of post-processing is that one does not need to be familiar with the local charac-

teristics of the weather quantity at hand, and can rely on its statistical model to fill in the gaps.

These statistical models or equations used in the post-processing methods, are actually mathematical representations of the relationship between the predictors and the predictands. To produce such formulas one should start with the data at hand and investigate the correlation between its variables. This has made the opportunity for new applications of these statistical models. In this thesis we present work that sought for novel probabilistic models for weather forecasting.

We use various statistical and machine learning methods for post-processing water levels at Kaub gauge in the Rhine River, solar irradiance at (3 and 7 locations in Germany and Hungary respectively) and total cloud cover of 3330 synoptic (SYNOP) stations around the world.

This thesis is organized as follows: First, in Chapter 1 we provide a literature review of post-processing, focusing on the models already used for the three weather quantities we study on. Then, in Chapter 2 we present the various post-processing methods used in this thesis. After that, in Chapter 3 we go through the evaluation methods used throughout followed by detailed case studies for each of the weather quantities, where we showcase the models, the data used and results (Chapters 4, 5 and 6). These chapters contain our new scientific results where we develop and implement new statistical post-processing methods, investigate the performance of the different models, develop and implement machine learning methods under various environments. The present results have been published in Baran et al. (2019); Schulz et al. (2021) and Baran et al. (2021). Finally, a chapter is dedicated to our conclusions summarising the results.

Chapter 1

Literature review of post-processing

1.1 Introduction

The prediction of every physical system is tainted by uncertainty, restricting our ability to forecast future states precisely. We make inherently imperfect mathematical models that describe the system's underlying rule in our endeavor to understand the behavior of a system. This uncertainty need to be quantified. In this chapter we are going to check the shift from point forecasts to probabilistic forecasts and the importance of post-processing. We will go through post-processing literature for the three weather quantities investigated in detail in Chapters 4, 5 and 6.

1.2 Probabilistic forecasting

If the time development of a system is sensitively dependent on the assumed initial conditions, it is said to display chaotic behavior. The system could be a mathematical system or a physical system, such as the Earth's atmosphere (such as a numerical model for weather forecasting). Because the exact state of the atmosphere can never be known, forecasts of its future evolution are always uncertain (White, 2007).

A NWP is a set of forecast values for different climatic variables at various

time intervals. These forecasts are made using a numerical model that solves a system of N differential equations resulting from a spatial discretization of the laws that control atmospheric dynamics (Buizza, 2018b). By its very nature, NWP is a subject that must cope with uncertainty. A NWP model's initial conditions can only be estimated to a certain degree of precision. Some of these initial errors (initial conditions uncertainty) can increase during a forecast, resulting in severe forecast errors. Furthermore, numerical algorithms' depiction of the dynamics and physics of the atmosphere includes additional uncertainties (model uncertainty), such as truncation errors and the uncertainty of parameters describing subgrid-scale processes like cumulus convection in a global model (Leutbecher and Palmer, 2008). Because the approximation of the initial conditions requires the use of a forecast model, initial condition uncertainties are influenced by model uncertainties.

In reality, meteorological centers compute the initial conditions based on a combination of observations and very short-term forecasts, often known as first guesses, utilizing statistical approaches. In the literature, this process is known as data assimilation (Buizza, 2018b).

The Ensemble of Data Assimilations (EDA) is a collection of data assimilations that takes into account uncertainty in observations, atmospheric boundary conditions including sea-surface temperature, and model physics (Lang et al., 2019). The EDA serves two purposes: it contributes to the high-resolution initial conditions (high-resolution analysis) by providing flow dependent estimates of the errors in the short-range forecasts (background) used in the data assimilation system; and it aids in the determination of the initial conditions for ensemble forecasts by providing EDA-based perturbations to the high-resolution analysis (Lang et al., 2019).

Other methods for quantifying uncertainty in NWP must deal with the sensitive dependency on initial conditions, the interaction of several spatial and temporal scales, and the reality that the sources of uncertainty are themselves uncertain (Leutbecher and Palmer, 2008). For example, some of the stochastic representations of these uncertainties at ECMWF are stochastically perturbed parametrization tendency (SPPT) and stochastic kinetic energy backscatter (SKEB) schemes (Leutbecher et al., 2017). Another approach to simulate initial uncertainties are singular vectors (SVs) and EDA-based perturbations (Buizza et al., 2010).

Moreover, probabilistic forecasting itself aims to measure uncertainty by making simple probability claims about the likelihood of future outcomes (Gneiting and Katzfuss, 2014). This dotes the start of shifting towards probabilistic forecasting (or ensemble forecasting). Since the climate system is dynamic

and it's behaviour is irregular, it cannot be predicted with confidence. Therefore probabilistic forecasts are preferable to deterministic forecasts.

The concept of ensemble forecasting in meteorology is introduced by Epstein (1969). Bauer et al. (2015) proposes a probabilistic method allowing representative sampling of the state of the atmospheric variable. In fact, a complete resolution of a probabilistic solution (by the Liouville and Fokker-Planck equations) is not possible from a computational point of view. The idea will be to slightly disturb the initial conditions and / or the physical parameterizations of the model resulting in not a single deterministic forecast but a set of forecasts (ensemble forecast members). For reasons of computation time, the members are produced at a spatial resolution which is generally less fine than the corresponding deterministic model.

The advent of high performance computing allowed the establishment of the first ensemble forecasting systems across the world (Toth and Kalnay, 1993; Mureau et al., 1993). Ensemble forecasting became a major tool for national meteorological services through the decision making support that it can provide to forecasters and the probability calculation of events of interest (Buizza et al., 1999; Palmer, 2002).

An ensemble forecasting system's principal goal is to evaluate forecast uncertainty by running a number of physically consistent future development scenarios. These scenarios are caused by uncertainty in the initial conditions as well as model error. The uncertainty in the boundary conditions is an additional source of forecast uncertainty for limited area ensembles (Reinert et al., 2021).

Effective alerts require accurate and informative weather forecasts to reduce both non-detections and false alarms. A good representation of the uncertainty is also necessary to have a reliable forecast of meteorological events. This makes it easier for users to make decisions, as they often have difficulty using probabilistic information (Hagedorn, 2017). Ensemble forecasting is more and more used in so-called weather-sensitive areas such as energy production (Taylor and Buizza, 2003; Pinson et al., 2009), hydrology (Krzysztofowicz, 2001; Schaake et al., 2007), agriculture (Calanca et al., 2010), ecology (Poulos et al., 2012), air quality (Mallet and Sportisse, 2006) or economy (Ravazzolo and Vahey, 2014).

1.3 The importance of statistical post-processing

Each ensemble forecast model has errors and biases that are not entirely random (there are often weaknesses due to the deterministic model that serves as the basis for generating the sets as well as a quantification of the erroneous uncertainty). Hence, to improve predictions, statistical post-processing methods have been used since the development of NWP (Glahn and Lowry, 1972).

The purpose of post-processing methods is to automatically build a statistical relationship between the observations and the corresponding meteorological variables predicted by the numerical model. Many data mining and machine learning techniques can be used (Hastie et al., 2009; Wu et al., 2014). These statistical models are then applied to the new forecasts in order to improve them, or to predict variables observed but not predicted by the numerical model. This approach, called ensemble Statistical Adaptation, predicts all or part of the distribution of the observation (Gneiting et al., 2005). A particularly interesting fact is that, whatever the initial performance of the ensemble forecasting model, a well-designed statistical post-processing manages to improve the performance of the forecasts (Ruth et al., 2009; Hemri et al., 2014; Taillardat et al., 2016).

The European Center for Medium-Range Weather Forecasts (ECMWF) has drawn up a report inventorying various existing statistical post-processing methods (Gneiting and Katzfuss, 2014). The Report of ECMWF discusses the potential operational application of state-of-the-art techniques in the field. Various methods of statistical calibration of ensemble forecasts for various weather variables have been established over the last decade, (see, e.g. Ruiz and Saulo, 2011; Schmeits and Kok, 2010; Williams et al., 2013; Wilks, 2018; Pinson and Messner, 2018), and parametric approaches such as ensemble model output statistics (EMOS; Gneiting et al., 2005) or Bayesian model averaging (BMA; Raftery et al., 2005) offer full predictive distributions. EMOS predictive distribution is given by a single parametric probability law with parameters depending on the ensemble, while BMA predictive probability density function (PDF) is a weighted mixture of PDFs corresponding to the individual ensemble members. In the applied parametric distribution family, the EMOS and BMA models differ in various weather quantities. Once a predictive distribution is given, its functionals (e.g. median or mean) can be considered as classical point forecasts. Non-parametric approaches like quantile regression (see e.g. Friederichs and Hense (2007); Bremnes (2019a)), and mixed methods like quantile mapping (see e.g. Hamill and Scheuerer (2018); Gascón et al. (2019)) also provide

estimates of the probability distributions of the weather quantities of interest.

Machine learning approaches also have recently become increasingly common in ensemble post-processing. For example, Taillardat et al. (2016) used quantile regression forests (QRF) for temperature and wind speed ensemble forecast calibration, and Taillardat et al. (2019) recently expanded the technique to precipitation forecasts. Rasp and Lerch (2018) used neural networks incorporating nonlinear relationship between predictor variables and forecast distribution parameters, while Bremnes (2019b) used neural networks in post-processing of ECMWF near-surface temperature ensemble forecasts using QRF as a benchmark model. Several machine learning methods, including random forests, gradient boosting, and neural networks, are compared by Bakker et al. (2019) for post-processing NWP predictions of solar irradiance based on quantile regression. For a detailed overview of new research trends and organizational implementations, see Vannitsem et al. (2021).

Probabilistic forecasting approaches that estimate predictive distributions are the most advanced prediction methods not only in the atmospheric sciences, but also in other fields of science and economy, such as economic risk management, seismic hazard prediction, and financial forecasting. For a detailed overview of the key concepts and properties of probabilistic predictions, as well as the areas of application see Gneiting and Katzfuss (2014).

In the following we will go through the post-processing methods used for the three investigated weather quantities.

1.4 Water level forecasting

In addition to the successful application, e.g. for temperature ensemble forecasts (Gneiting et al., 2005), wind speed (Baran and Lerch, 2015; Lerch and Thorarinsdottir, 2013; Thorarinsdottir and Gneiting, 2010) or precipitation (Baran and Nemoda, 2016; Scheuerer, 2013; Scheuerer and Hamill, 2015), EMOS-based statistical post-processing has been shown to improve the predictive performance of hydrological ensemble forecasts for different gauges along the Rhine river (Hemri et al., 2015; Hemri and Klein, 2017), too EMOS is a fairly parsimonious post-processing approach, and its performance is then restricted by i) the degree to which the true mechanism can be interpreted by the parametric distribution family and ii) the extent to which the complete information from the ensemble can be summarized in a limited set of ensemble statistics.

For example, a typical EMOS approach based on a Gaussian or Gamma distribution family is not capable of modeling bimodal forecast distributions.

However, BMA, which has also been applied to hydrological ensemble forecasts (Duan et al., 2007; Hemri et al., 2013), is more flexible in that it converts a (multi-model) raw ensemble into a mixture distribution that allows multimodal shapes. Accordingly, in this thesis, we assume that BMA may be able to outperform EMOS.

1.5 Solar irradiance forecasting

For decades, the literature on energy forecasting has largely concentrated on deterministic predictions. However, it is now generally accepted that probabilistic forecasting is critical for making the best decisions in planning and operating solar power (Hong and Fan, 2016; Van der Meer et al., 2018; Haupt et al., 2019; Hong et al., 2020). In their recent analysis on energy forecasting, Hong et al. (2020) identified probabilistic forecasting with the aim of providing a predictive probability distribution for a future quantity or event in order to measure forecast uncertainty (variance, probabilities of different events, etc.) as one of the most relevant emerging research topics.

We will concentrate on solar energy, which is one of the most effective renewable energy sources connected to photovaltaic (PV) power. Photovoltaics is the process of converting sunlight into electricity using semi-conductors and for example, PV contributes significantly to Germany's power supply, accounting for 9.2% of gross electricity consumption in 2020 and more than 66% of current electricity consumption on sunny days, according to Fraunhofer Institute for Solar Energy Systems (2021).

There are two types of solar energy forecasting approaches: those that aim to predict solar irradiance and those that aim to predict PV power. Solar irradiance and PV system output are naturally highly correlated, and the statistical methods used are similar (Van der Meer et al., 2018). In this thesis, we'll concentrate on probabilistic solar irradiance forecasting, see Van der Meer et al. (2018) and Yang (2019) for detailed overviews and analyses of current approaches. Most methods of the ones for probabilistic solar irradiance forecasting combine physical knowledge from NWP models with statistical methods, with the exception of short-term prediction (e.g., Zelikman et al., 2020).

In the literature on probabilistic solar energy forecasting, post-processing ensemble predictions of solar irradiance has recently got a lot of attention. It should be noted that similar post-processing methods have also been used for direct PV power production forecasting, such as in Sperati et al. (2016), but they are out of scope in this thesis. Bakker et al. (2019) compares clear-sky in-

dex post-processing methods that use deterministic NWP predictions of many variables as input and use different machine learning approaches for quantile regression. Le Gal La Salle et al. (2020) proposes an EMOS model for global horizontal irradiance, comparing quantile regression and analog ensemble approaches to truncated normal and generalized extreme value distributions as forecast laws. For hourly clear-sky index forecasting, Yagli et al. (2020) compares several parametric and nonparametric post-processing methods, including EMOS models based on Gaussian, truncated logistic, and skewed Student's t distributions, quantile regression based on random forests, and generalized additive models for location, scale, and shape. Yang (2020a) suggests the use of EMOS models for probabilistic site adaptation of gridded solar irradiance products, and Yang (2020b) compares building models for irradiance and the clear-sky index, and examines the option of parametric distributions in a closely related paper.

1.6 Total Cloud Cover forecasting

According to the World Meteorological Organization, "Total cloud cover is the proportion of the sky covered by all visible clouds" (World Meteorological Organization, 2017). Total cloud cover (TCC) observations are typically recorded in eighths of sky cover called oktas, which take only nine different values, despite the fact that the definition implies a continuous quantity in the $[0, 1]$ interval.

TCC forecasts are produced using numerical NWP models (see Køltzow et al. (2019) for a comparison of the performance of state-of-the-art techniques), and all major meteorological centers have recently issued ensemble TCC forecasts using their operational ensemble prediction systems (EPSs). The EPS of the independent intergovernmental ECMWF (Molteni et al., 1996; Leutbecher and Palmer, 2008; ECMWF Directorate, 2012) or the Global Ensemble Forecast System of National Centers for Environmental Prediction (Zhou et al., 2017) are good instances.

TCC ensemble forecasts are typically highly underdispersive and have systemic bias, they underperform ensemble forecasts of other weather variables such as temperature, wind speed, pressure, or precipitation in terms of forecast skill (see, for example, Haiden et al. (2015, 2018)).

Since TCC is discrete, the predictive distribution should be a discrete probability distribution, and post-processing can be thought of as a classification problem that yields the oktas' probabilities. Hemri et al. (2016) propose two discrete parametric post-processing approaches for calibrating TCC ensemble

forecasts: multiclass logistic regression (MLR) (Izenman, 2008) and proportional odds (or ordered) logistic regression (POLR) (McCullagh, 1980). Various versions of logistic regression had already been successfully implemented in statistical post-processing (see, for example, Wilks (2009); Schmeits and Kok (2010)), and ordered logistic regression even performed well for forecasting discrete categories (Messner et al., 2014).

Chapter 2

Post-processing approaches

2.1 Introduction

In this chapter, we are going to go through all post-processing approaches used in this thesis. Classical parametric approaches are applied in water level and solar irradiance forecasting, whereas for total cloud cover we use machine learning techniques.

2.2 Statistical approaches

In what follows, let $f_1, \ldots, f_K$ denote the ensemble forecast of a given weather quantity X for a given location, time and lead time. However, most operational EPSs now include ensembles in which at least some of the members are statistically indistinguishable, therefore they are considered exchangeable. As an example one can consider the 51-member operational EPS of the ECMWF or the 11-member AROME-EPS of the Hungarian Meteorological Service (HMS).

In the case of exchangeability, having M ensemble members that are divided into K exchangeable groups where the kth group contains $M_k \geq 1$ $\left(\sum_{k=1}^{K} M_k = M\right)$ forecasts, we denote the ℓth member of the kth group by $f_{k,\ell}$.

2.2.1 Bayesian model averaging

For the BMA post-processing approach, the predictive distribution of a future weather quantity is a weighted mixture of probability laws corresponding to the

individual ensemble members. The BMA predictive PDF (Raftery et al., 2005) of X equals

$$p(x|\, f_1, \ldots, f_K; \theta_1, \ldots, \theta_K) := \sum_{k=1}^{K} \omega_k g\big(x|\, f_k, \theta_k\big), \tag{2.1}$$

where $g\big(x|f_k, \theta_k\big)$ is the component PDF from a parametric family corresponding to the kth ensemble member f_k with parameter (vector) θ_k to be estimated, and ω_k is the corresponding weight linked to the relative performance of this particular member during the training period. Note that the weights should form a probability distribution, that is $\omega_k \geq 0,\ k = 1, \ldots, K$, and $\sum_{k=1}^{K} \omega_k = 1$.

Fraley et al. (2010) propose using the same weights and parameters inside a particular group to account for the presence of groups of exchangeable ensemble members. Therefore, in this situation, equation (2.1) is replaced by

$$p\big(x|f_{1,1}, \ldots, f_{1,M_1}, \ldots, f_{K,1}, \ldots, f_{K,M_K}; \theta_1, \ldots, \theta_K\big) := \sum_{k=1}^{K} \sum_{\ell=1}^{M_k} \omega_k g\big(x|f_{k,\ell}, \theta_k\big) \tag{2.2}$$

To model temperature and sea level pressure for example, Raftery et al. (2005) proposed a normal mixture with the following predictive distribution

$$\sum_{k=1}^{K} \omega_k \mathcal{N}\big(a_k + b_k f_k, \sigma^2\big), \tag{2.3}$$

where $\mathcal{N}\big(\mu, \sigma^2\big)$ denotes a Gaussian distribution with mean μ and variance σ^2.

2.2.2 Ensemble model output statistics

The EMOS or non-homogeneous regression approach, proposed by Gneiting et al. (2005), uses a single parametric distribution as a predictive PDF with parameters linked to the ensemble members. Those parameters depend on the kernel used in the EMOS model which has the following general form

$$X|f_1, \ldots, f_K \sim h\big(x|f_1, \ldots, f_K; \theta\big) \tag{2.4}$$

where $h\big(x|f_1, \ldots, f_K; \theta\big)$ is a parametric PDF.

For example, in modeling temperature and sea level pressure, Gneiting et al. (2005) proposed the following EMOS predictive distribution

$$\mathcal{N}\big(a_0 + a_1 f_1 + \ldots + a_k f_k, b_0 + b_1 S^2\big) \quad \text{with} \quad S^2 := \frac{1}{K-1} \sum_{k=1}^{K} \big(f_k - \overline{f}\big) \tag{2.5}$$

where $\overline{f}$ is the ensemble mean.

In case of exchangeability, the model (2.5) becomes

$$\mathcal{N}\big(a_0 + a_1 \overline{f_1} + \ldots + a_K \overline{f_K}, b_0 + b_1 S^2\big)$$

where $\overline{f_k}$ denotes the mean of the kth exchangeable group, $k = 1, 2, \ldots, K$.

2.3 Post-processing methods for discrete quantities

Let $Y \in \mathcal{Y} = \{y_1, y_2, \ldots, y_n\}$ be a discrete weather quantity observed at a given location and time point. A good example is TCC to be discussed in detail in Chapter 6, which is expressed in oktas, that is it can take $n = 9$ different values. In this case post-processing is reduced to a classification problem where the aim is to predict the probabilities of the different outcomes based on feature vectors $\boldsymbol{x}$ derived using ensemble members or/and other covariates. Thus, one can use either classical parametric approaches to classification like MLR or POLR to be discussed in details in Sections 2.3.1 and 2.3.2, respectively, or apply some state-of-the-art machine learning classification method (see Section 2.4).

2.3.1 Multiclass logistic regression

In MLR the log-odds of a given class with respect to a fixed reference class, which can be chosen arbitrarily, is represented as an affine function of the features. This means that after setting e.g. the last class y_n as reference class, the conditional distribution of the predicted weather quantity with respect to an M-dimensional feature vector $\boldsymbol{x}$ equals

$$\mathsf{P}\big(Y = y_k \mid \boldsymbol{x}\big) = \begin{cases} \frac{\mathrm{e}^{L_k(\boldsymbol{x})}}{1+\sum_{\ell=1}^{n-1} \mathrm{e}^{L_\ell(\boldsymbol{x})}}, & k = 1, 2, \ldots, n-1; \\ \frac{1}{1+\sum_{\ell=1}^{n-1} \mathrm{e}^{L_\ell(\boldsymbol{x})}}, & k = n, \end{cases} \quad \text{with} \tag{2.6}$$

$$L_k(\boldsymbol{x}) := \beta_{0k} + \boldsymbol{x}^\top \boldsymbol{\beta}_k,$$

where $\beta_{0k} \in \mathbb{R}$, $\boldsymbol{\beta}_k \in \mathbb{R}^M$, resulting in $(n-1)(M+1)$ free parameters to be estimated on the basis of the training data.

2.3.2 Proportional odds logistic regression

The POLR model is designed to fit ordered data. Given a feature vector $\boldsymbol{x}$, the conditional cumulative probabilities of Y are expressed as

$$\mathsf{P}\big(Y \leq y_k \mid \boldsymbol{x}\big) = \frac{\mathrm{e}^{\mathcal{L}_k(\boldsymbol{x})}}{1+\mathrm{e}^{\mathcal{L}_k(\boldsymbol{x})}}, \quad \text{with} \quad \mathcal{L}_k(\boldsymbol{x}) := \gamma_{0k} + \boldsymbol{x}^\top \boldsymbol{\gamma}, \quad k = 1, 2, \ldots, n, \quad (2.7)$$

where we assume that $\gamma_{01} < \gamma_{02} < \cdots < \gamma_{0n}$. In this way POLR model (2.7) is more parsimonious than MLR model (2.6), as it has just $n + M$ unknown parameters.

2.4 Machine learning approaches

In this thesis, the machine learning approaches were only used for the TCC weather quantity, where the feature vector $\boldsymbol{x}$ is derived from the ensemble members or/and other covariates.

2.4.1 Multilayer perceptron neural network

A multilayer perceptron (MLP) is a type of feedforward neural network that consists of an input layer, an output layer, and several intermediate layers (also known as hidden layers) that each contain several neurons. The value in each of the neurons is a transformed value (via an activation function) of a weighted sum of all neuron values from the previous layer plus a bias term. The number of neurons in the input and output layers is determined by the number of features and classes, respectively, while the number of hidden layers and the number of neurons in each hidden layer are network parameters that are free (or tuning). For a detailed introduction to neural networks, see, for example Goodfellow et al. (2016).

The weights of the neurons are calculated in order to minimize a given loss function on the training set, and the network is trained using a set of labeled data (training set). Early stopping rules based on a validation set are recommended to prevent overfitting. Typically, a subset of the labelled data set available for training is selected randomly. If the value of the loss function computed on the

validation set does not change after a certain number of iterations, the minimization process ends. Other machine learning approaches use similar techniques; see Section 6.4.1 for more details.

The addition of a regularization term to the loss function is another way to avoid overfitting. We use a L_2 regularization here, in which the sum of squares of the network's weights is multiplied by a factor (which is an additional tuning parameter of the network). A probability distribution corresponding the different classes (oktas for TCC) provided by the trained network for each feature vector.

2.4.2 Random forest models and gradient boosting machines

Machine learning models based on ensembles of decision trees include random forests (RF) and gradient boosting machines (GBM). Since the 1950s, decision trees have been used in meteorological forecasting as flowchart-like structures (McGovern et al., 2017). These models are obtained by iteratively splitting training data into groups according to a threshold in one of the features $\boldsymbol{x}$ which is chosen to maximize the homogeneity of the target variable within the resulting subsets. This process is repeated until a criterion for stopping is met. Out-of-sample forecasts can be obtained by recursively partitioning the feature space according to the predictor input and estimating class probabilities using the empirical frequencies of observed classes in the corresponding subset. Although there are multiple decision tree learning algorithms, we will concentrate in this thesis on classification and regression trees (CART), which were first introduced by Breiman et al. (1984).

Random forest models

RF models (Breiman, 2001) repeatedly resample the training set to obtain multiple decision trees to increase robustness and solve overfitting issues of decision trees. At each splitting node, this bootstrap aggregation (or bagging) method is combined with only considering a random subset of the predictors. Averaging over the decision trees in the RF ensemble yields class probability predictions for out-of-sample instances.

When implementing RF models, several tuning parameters must be selected. Most significantly, the number of trees in the forest must be determined, as well as the depth (number of levels of recursive partitioning) and number of predictor variables randomly chosen at each splitting node for each tree. RF

models, in general, are relatively resistant to these tuning parameters and are less susceptible to overfitting for a wide variety of parameter choices.

Gradient boosting machines

GBM, as opposed to randomly resampling the training data, are made up of ensembles of decision trees that are grown sequentially using data from previously grown trees. As a result, each decision tree is fitted to a changed version of the original training set, with a focus on areas where previous model iterations failed to predict accurately.

The term "boosting" refers to a class of machine learning algorithms that suit models by combining many simpler models, in this thesis we used decision trees. Various notions of gradient boosting have been established following Friedman (2000), and it has been demonstrated that boosting can be viewed as a gradient descent algorithm in function space where a loss function is iteratively optimized by selecting a function that points in the direction of the negative gradient. Gradient boosting principles can be applied to a wide range of loss functions, and algorithms for a variety of machine learning tasks have been developed, see, for example, Hastie et al. (2009) for a general introduction to gradient boosting.

We use extreme gradient boosting (Chen and Guestrin, 2016), a form of tree-based gradient boosting that relies on second-order approximations of the objective function. GBM model predictions are obtained via

$$\hat{z}^c = \sum_{m=1}^{M} h_m^c(\boldsymbol{x}), \tag{2.8}$$

where h_m^c denotes a regression tree for category $c \in \{1, \ldots, n\}$ containing a continuous value in all terminal leaves, and M is the number of boosting iterations. For probabilistic classification tasks, separate sets of regression trees are fitted simultaneously for all categories, and the obtained latent values $\hat{z}^c$ are transformed according to a softmax function. A regularized version of the LogS (see Section 3.2) is used to learn the set of functions used in the model (2.8), for details, see Chen and Guestrin (2016).

GBM often outperform RF models in a number of applications, but they are more susceptible to overfitting and more difficult to tune than RF models. The number of boosting iterations, M, is particularly important. Furthermore, the complexity of individual trees h_m must often be limited; see Section 6.4.1 for more details.

2.5 Estimation of model parameters and training data selection

Typically, model parameters are estimated using training data, comprised of ensemble members and verifying observations of previous dates. The most common one is rolling training period (RTP), where for a certain verification date, we use the previous n days. Smaller datasets may benefit from RTP, which enable models to adjust to changes in meteorological conditions or the underlying NWP system. Another option is to use all available data by considering longer training periods, as shown by studies that demonstrate using long archives of training data, regardless of possible NWP model changes during that period, often result in superior performance (Lang et al., 2020). In operational implementations, extending training periods on a regular basis may also be useful, where data archives are built up and extended over time.

There are two traditional techniques to spatial selection: local and regional (often sometimes referred to as global) (Thorarinsdottir and Gneiting, 2010). In the local approach, parameters for a specific location are computed using just data from that location, resulting in different parameter estimates for different locations. Local modeling needs long time periods for training in order to ensure numerical stability of the estimation process, which is the main drawback of this method. The regional selection, on the other hand, uses training data from the whole ensemble domain and the same set of parameters for all locations. Since local modeling adresses location-specific forecast error characteristics, it often produces better forecast skill than the regional one. Examples of all mentioned training data selection methods are seen in the case studies of Section 5.4: regional estimation with a rolling training period in Section 5.4.1 and local estimation with rolling and extending training periods in Section 5.4.2.

In this thesis we used other methods for training data selection, e.g. for a certain verification date we select data only of the same season from the previous years. Other methods such as analog-based ones, detailed in Section 4.2.5.

Chapter 3

Forecast evaluation

3.1 Introduction

The main aim of probabilistic forecasting, as mentioned in Gneiting et al. (2007), is to maximize the sharpness of the predictive distribution subject to calibration. Sharpness refers to the predictive distribution's concentration (see e.g. Lauret et al. (2019)), while calibration refers to the statistical consistency between forecasts and observations. These two goals can be met at the same time by using proper scoring rules, which are loss functions $\mathcal{S}(F, x)$ that assign numerical values to pairs (F, x) of forecasts and observations. For each case of weather quantity, we use several scores presented in this chapter.

3.2 Logarithmic score

The logarithmic score of a continuous distribution F is defined as

$$\text{LogS}\left(F, x\right) := -\log\left(f(x)\right),$$

where f is the density corresponding to the Cumulative Distribution Function (CDF) F, whereas in the discrete case, the logarithmic score is the negative logarithm of the probability mass function (PMF) evaluated at the observation, that is

$$\text{LogS}\left(F, x\right) := -\log\left(p_F(x)\right).$$

LogS is proper and negatively oriented, that is the smaller values indicate better forecast.

3.3 Continuous ranked probability score

In atmospheric sciences, one of the most popular scoring rules is the continuous ranked probability score (CRPS; Gneiting and Raftery, 2007; Wilks, 2019), as it assesses calibration and sharpness simultaneously. For a predictive CDF F and real-valued observation x, the CRPS is defined as

$$\mathrm{CRPS}\left(F, x\right) := \int_{-\infty}^{\infty} \left(F(y) - \mathbb{1}_{\{y \geq x\}}\right)^2 \mathrm{d}y = \int_{-\infty}^{x} F^2(y) \mathrm{d}y + \int_{x}^{\infty} \left(1 - F(y)\right)^2 \mathrm{d}y \tag{3.1}$$

$$= \mathsf{E}|X - x| - \frac{1}{2}\mathsf{E}|X - X'|,$$

where $\mathbb{1}_H$ denotes the indicator of a set H, while X and X' are independent random variables with CDF F and finite first moment. The CRPS is also a negatively oriented score, and the third representation in (3.1) implies that it can be expressed in the same unit as the observation. Further, it can be seen as a generalization of the absolute error to probabilistic forecasts and allows comparisons with point forecasts. The CRPS of an ensemble forecast is defined with the help of the empirical CDF.

Discrete Case

The CRPS for the discrete case can be given as

$$\mathrm{CRPS}\left(F, x\right) = \sum_{k=1}^{n} p_F(y_k)\left|y_k - x\right| - \sum_{k=2}^{n} \sum_{\ell=1}^{k-1} p_F(y_k) p_F(y_\ell)\left|y_k - y_\ell\right|.$$

3.4 Improvements in scores

For a given probabilistic forecast F, the improvement in a score $\mathcal{S}_F$ with respect to a reference forecast F_{ref} can be quantified with the help of the corresponding skill score defined as

$$\mathcal{SS}_F := 1 - \frac{\overline{\mathcal{S}}_F}{\overline{\mathcal{S}}_{F_{\mathrm{ref}}}},$$

where $\overline{\mathcal{S}}_F$ and $\overline{\mathcal{S}}_{F_{\mathrm{ref}}}$ denote the mean score values over the verification data corresponding to forecasts F and F_{ref}, respectively.

In particular, over all forecast cases in the verification data, the goodness of fit of competing forecasts in terms of probability distributions is compared with the aid of the mean CRPS and mean LogS values $\overline{\text{CRPS}}$ and $\overline{\text{LogS}}$, respectively. Further, the improvement in CRPS and LogS with respect to a reference model can be quantified using the continuous ranked probability skill score (CRPSS) and logarithmic skill score (LogSS), respectively, defined as

$$\text{CRPSS} := 1 - \frac{\overline{\text{CRPS}}}{\overline{\text{CRPS}}_{ref}} \qquad \text{and} \qquad \text{LogSS} := 1 - \frac{\overline{\text{LogS}}}{\overline{\text{LogS}}_{ref}},$$

where $\overline{\text{CRPS}}_{ref}$ and $\overline{\text{LogS}}_{ref}$ denote the mean CRPS and mean LogS of the reference approach (see e.g. Gneiting and Raftery (2007); Murphy (1973)). It's worth noting that both CRPSS and LogSS are positively oriented, meaning that higher skill scores indicate better predictive performance.

3.5 Brier score

For assessing the predictive performance of the different forecasts with respect to the binary event that observation x exceeds a given threshold y, one can consider the Brier score (BS; Wilks, 2019, Section 9.4.2), which for a predictive CDF F is defined as

$$\text{BS}\left(F, x; y\right) := \left(F(y) - \mathbb{1}_{\{y \geq x\}}\right)^2 \tag{3.2}$$

(see e.g. Gneiting and Ranjan, 2011). Note that the BS is also negatively oriented and the CRPS is the integral of the BS over all possible thresholds.

3.6 Probability integral transform

One of the simplest tools for getting a first impression about the calibration of forecast distributions is the probability integral transform (PIT) histogram. By definition, the PIT is the value of predictive CDF at the validating observation (Raftery et al., 2005), with possible randomization at points of discontinuity (Gneiting and Ranjan, 2013). In case of proper calibration, PIT should follow a uniform distribution on the $[0, 1]$ interval. In the case where uniformity is not achieved, the shape of the PIT histogram provides information about the possible cause of the problem. In this way the PIT histogram is the continuous counterpart of the verification rank histogram for the raw ensemble, which is

defined as histogram of ranks of validating observations with respect to the corresponding ensemble forecasts (see e.g. Wilks, 2019, Section 7.7.2). Again, for a properly calibrated ensemble the ranks should be uniformly distributed.

3.7 Coverage and average width

Calibration and sharpness of a predictive distribution can also be investigated using the coverage and average width of the $(1-\alpha)100\,\%$, $\alpha \in (0,1)$, central prediction interval, respectively (Gneiting et al., 2007). As coverage we consider the proportion of validating observations located between the lower and upper $\alpha/2$ quantiles of the predictive CDF, and level α should be chosen to match the nominal coverage of the raw ensemble, that is $(K-1)/(K+1)100\%$, where K is the ensemble size (see e.g. Raftery et al., 2005; Baran and Lerch, 2015). As the coverage of a calibrated predictive distribution should be around $(1-\alpha)100\,\%$, such a choice of α allows direct comparison with the raw forecast.

3.8 MAE and RMSE

Point forecasts such as median and mean of the raw ensemble and of the predictive distribution are evaluated with the help of the mean absolute error (MAE) and root mean square error (RMSE). Note that the former is optimal for the median, whereas the latter for the mean (Gneiting, 2011).

3.9 Diebold-Mariano test

As suggested by Gneiting and Ranjan (2011), statistical significance of the differences between the verification scores is assessed by utilizing the Diebold-Mariano test (DM; Diebold and Mariano, 1995), which allows accounting for the temporal dependencies in the forecast errors. The detailed description of the DM test can be found for example in Baran and Lerch (2016).

Chapter 4

Post-processing of probabilistic water level forecasts

4.1 Introduction

Hydrological forecasts are essential for a diverse group of users, such as operators of hydrological power plants, flood prevention authorities or shipping companies. Any forecasting should be provided for rational decision-making based on cost-benefit analysis of the predictive uncertainty (Krzysztofowicz, 1999; Todini, 2008). The state-of-the-art method of using a series of parallel runs of a hydrological models driven by the meteorological ensemble forecast given by the NWPs. Cloke and Pappenberger (2009) offers a first estimate of the meteorological input uncertainty. However, NWP ensembles are usually biased and underdispersed (Bougeault et al., 2010; Buizza et al., 2005; Park et al., 2008). Additional sources of uncertainty, such as hydrological model formulation, boundary and initial uncertainty as well as measurement uncertainties, are usually neglected. Statistical post-processing is therefore necessary in order to minimize systematic errors and to obtain an accurate estimation of the predictive uncertainty (Buizza, 2018a).

BMA models with Gaussian components offer a suitable fit for weather vari-

ables such as temperature or pressure (Fraley et al., 2010; Raftery et al., 2005), whereas wind speed requires a non-negative and skewed distribution such as gamma (Sloughter et al., 2010) or a truncated normal distribution from below at zero (Baran, 2014). Although water levels are usually non-Gaussian (see, e.g. Duan et al., 2007), it is true that they are bounded both from above and below. It is imperative to also account for these constraints when formulating the model.

Since the use of the BMA approach is very convenient in the Gaussian framework, both Duan et al. (2007) and Hemri et al. (2013) perform a Box-Cox transformation (4.1) before applying BMA in order to achieve approximate normality despite the positive skewness of water level. In addition, it is necessary to ensure that the water level quantiles resulting from the predictive distribution are within the practical physical bounds. At the upper bound of the distribution level, the water level should be less than level threshold resulting from an extreme flood with a slight overflow probability, and at the lower bound water levels should be greater than a threshold resulting from an extreme long-lasting low water duration with a small non-exceeding probability. In order to ensure practical values while also being able to benefit from the mathematical simplicity of Gaussian models, a lower and upper truncated normal distribution is used. The data of the water level gauges is usually established in order to prevent water levels below zero. As a consequence, we use an ad hoc lower truncation limit of half the lowest value ever registered. While this ad hoc limit turned out to be appropriate for the gauge considered, it is usually suggested that the derivation of the lower truncation limit be based on physical properties such as the cease-to-flow stage. While river physics does not provide a hard upper limit for the water level, as a result of the Box-Cox transformation we need to apply an upper limit for the truncation. Otherwise, due to the skewness of the water level, the back-transformation of the Box-Cox transformed into the original space can lead to predicted water levels that are far above the rating curve range. In this chapter, the upper truncation limit has been set to two times the highest gauge level ever reported. However, in general, we suggest that the upper limit be based on an evaluation of the range of validity of the rating curve. Above this range, the hydrometric gauge may be bypassed and, thus, the rating curve becomes obsolete.

To our best knowledge, up to now, there is no study that has applied a doubly truncated normal BMA approach. In this chapter, the work of Baran (2014), which applies a one-sided truncated normal BMA method, is adapted to a two-sided truncated normal BMA approach. Its performance and

suitability for hydrological ensemble forecasting is tested by using the example of a multi-model ensemble forecasts of water level at the Kaub gauge in the Rhine River.

Doubly truncated BMA is introduced in Section 4.2, followed by a brief description of the data in Section 4.3. The results are presented in Section 4.4, see also Baran et al. (2019).

4.2 Post-processing methods

4.2.1 Truncated Normal BMA model

One standard practice is to normalize forecasts and observations, for example, by applying the Box-Cox transformation

$$h_\lambda(x) := \begin{cases} \big(x^\lambda - 1\big)/\lambda, & \lambda \neq 0, \\ \log(x), & \lambda = 0 \end{cases} \tag{4.1}$$

with some coefficient λ, perform post-processing, and then back transform the results using the inverse Box-Cox transformation (Duan et al., 2007; Hemri et al., 2014, 2015). Following the ideas of Hemri and Klein (2017), to model Box-Cox transformed water levels, we use a doubly truncated normal distribution $\mathcal{N}_a^b\big(\mu, \sigma^2\big)$, with PDF

$$g_{a,b}\big(x|\,\mu,\sigma\big) := \frac{\frac{1}{\sigma}\varphi\big(\frac{x-\mu}{\sigma}\big)}{\Phi\big(\frac{b-\mu}{\sigma}\big) - \Phi\big(\frac{a-\mu}{\sigma}\big)}, \quad x \in [a,b], \tag{4.2}$$

and $g_{a,b}\big(x|\,\mu,\sigma\big) := 0$ otherwise, where a and b are the lower and upper bounds and φ and Φ are the PDF and CDF of the standard normal distribution, respectively. Note, that the mean and variance of $\mathcal{N}_a^b\big(\mu, \sigma^2\big)$ are

$$\begin{aligned} \kappa &= \mu + \sigma \frac{\varphi\big(\frac{a-\mu}{\sigma}\big) - \varphi\big(\frac{b-\mu}{\sigma}\big)}{\Phi\big(\frac{b-\mu}{\sigma}\big) - \Phi\big(\frac{a-\mu}{\sigma}\big)} \qquad \text{and} \\ \varrho^2 &= \sigma^2 \left(1 + \frac{\frac{a-\mu}{\sigma}\varphi\big(\frac{a-\mu}{\sigma}\big) - \frac{b-\mu}{\sigma}\varphi\big(\frac{b-\mu}{\sigma}\big)}{\Phi\big(\frac{b-\mu}{\sigma}\big) - \Phi\big(\frac{a-\mu}{\sigma}\big)} - \left(\frac{\varphi\big(\frac{a-\mu}{\sigma}\big) - \varphi\big(\frac{b-\mu}{\sigma}\big)}{\Phi\big(\frac{b-\mu}{\sigma}\big) - \Phi\big(\frac{a-\mu}{\sigma}\big)}\right)^2\right), \end{aligned} \tag{4.3}$$

respectively. The proposed BMA predictive PDF is defined as

$$p\big(x|\, f_1,\ldots,f_K;\alpha_1,\ldots,\alpha_K;\beta_1,\ldots,\beta_K;\sigma\big)=\sum_{k=1}^{K}\omega_k g_{a,b}\big(x|\,\alpha_k+\beta_k f_k,\sigma\big),\quad(4.4)$$

where we assume that the location of the kth mixture component is an affine function of the corresponding f_k ensemble member and the scale parameters are assumed to be equal for all PDF components. The last assumption is for simplicity's sake and is common in the BMA modeling (Raftery et al., 2005, see, e.g.), while the form of location parameter, is consistent with the truncated normal BMA model of Baran (2014). The estimation of model paramters is detailed in Section 4.2.2.

An example of the BMA predictive distribution (4.4) of Box-Cox transformed water levels for 30 July 2008 is shown in Figure 4.1, where the overall predictive PDFs and the component PDF are plotted together with the BMA median and the transformed validating observation.

Note that, a standard EMOS approach based on a Gaussian or Gamma distribution families is incapable of modeling bimodal forecast distributions. BMA, on the other hand, is more flexible as it transforms a (multimodel) raw ensemble to a mixture distribution, allowing for multimodal shapes. This justifies the use of BMA in post-processing.

4.2.2 Parameter estimation

Location parameters α_k, β_k, weights ω_k, $k=1,\ldots,K$; and scale parameter σ of the Truncated Normal BMA model (4.4) are estimated from training data, which comprises of ensemble members and corresponding validating observations for a given time period of length n. Usually, for the BMA approach, location parameters are calculated by regressing validating observations on the ensemble members; while, scale and weight parameters are estimated using the EM algorithm for mixture distribution (Dempster et al., 1977; McLachlan and Krishnan, 1997) to optimize (maximize) the log likelihood function (Raftery et al., 2005; Sloughter et al., 2007, 2010) of the training data. Although the location parameters are considered to be simple functions of the mean, the regression method is not really suitable as the mean does not equal the location. Therefore we propose a pure ML method estimating all model parameters by maximizing the likelihood function, which ideas have already been considered, for example, by Sloughter et al. (2010); Baran (2014).

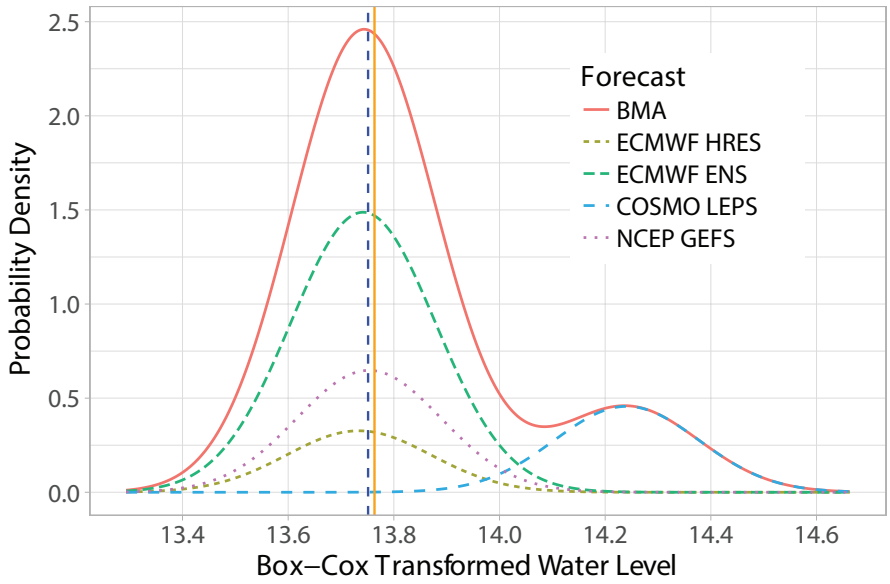

Figure 4.1: BMA predictive distribution and its components of Box-Cox transformed water levels for 30 July 2008 corresponding to 48 hours lead time. Vertical solid line: transformed verifying observation; dashedline: BMA median.

Let $f_{k,s,t}$ denote the kth ensemble member, and let $x_{s,t}$ denote the corresponding validating observation for a location $s \in \mathcal{S}$ and time $t \in \mathcal{T}$, where $\mathcal{S}$ denotes the set of locations sharing the same BMA model parameters and $\mathcal{T}$ denotes the set of training dates. For the case study of Section 4.4, $\mathcal{S}$ consists of a single location; however, for more complex ensemble domains, different choices of training data are possible, as described in Lerch and Baran (2016). In Section 4.4, the different lead times are addressed separately, and thus the lead time of the forecast is omitted. Assuming conditional independence of forecast errors in space and time, the log-likelihood function for model (4.4) for all forecast cases

(s,t) in the training set is as follows

$$\ell(\omega_1,\ldots,\omega_K,\alpha_1,\ldots,\alpha_K,\beta_1,\ldots,\beta_K,\sigma) \tag{4.5}$$
$$= \sum_{s,t} \log\left[\sum_{k=1}^{K} \omega_k g_{a,b}\big(x_{s,t}|\,\alpha_k+\beta_k f_{k,s,t},\sigma\big)\right].$$

We apply the EM algorithm for truncated Gaussian mixtures suggested by Lee and Scott (2012) and add a mean correction to obtain the ML estimates. In line with the classic EM algorithm for mixtures, we first implement latent binary indicator variables $z_{k,s,t}$ which classify the mixture component from which observation $x_{s,t}$ originates, i.e. $z_{k,s,t}$ is one or zero based on whether or not $x_{s,t}$ follows the distribution of the kth component. Using these indicator variables, the full data log-likelihood corresponding to (4.5) can be given in the following form.

$$\ell_C(\omega_1,\ldots,\omega_K,\alpha_1,\ldots,\alpha_K,\beta_1,\ldots,\beta_K,\sigma) \tag{4.6}$$
$$= \sum_{s,t}\sum_{k=1}^{K} z_{k,s,t}\left[\log\left(\omega_k\right)+\log\Big(g_{a,b}\big(x_{s,t}|\,\mu_{k,s,t},\sigma\big)\Big)\right],$$

with $\mu_{k,s,t} := \alpha_k+\beta_k f_{k,s,t}$. After defining the initial values of the parameters, the EM algorithm alternates between the expectation (E) and the maximization (M) step up to the convergence. As first guesses $a_k^{(0)}$ and $b_k^{(0)}$, $k=1,\ldots,K$, for location parameters we consider the linear regression coefficients of $x_{s,t}$ on $f_{k,s,t}$, so $\mu_{k,s,t}^{(0)} = \alpha_k^{(0)}+\beta_k^{(0)} f_{k,s,t}$. Initial scale $\sigma^{(0)}$ may be the standard deviation of the observations in the training data set or the average residual standard deviation from the latter regression, while the original weights may be chosen uniformly, i.e. $\omega_k^{(0)} = 1/K$, $k=1,\ldots,K$. Then, in the E step, the latent variables are estimated using the conditional expectation of the complete log-likelihood on the observed data, while in the M step, the estimations of the parameters are updated by maximizing ℓ_C given the actual values of the latent variables. For the doubly truncated normal model specified by (4.2) and (2.1), the E step of the $(j+1)$st iteration is

$$z_{k,s,t}^{(j+1)} := \frac{\omega_k^{(j)} g_{a,b}\big(x_{s,t}|\,\mu_{k,s,t}^{(j)},\sigma^{(j)}\big)}{\sum_{i=1}^{K}\omega_i^{(j)} g_{a,b}\big(x_{s,t}|\,\mu_{i,s,t}^{(j)},\sigma^{(j)}\big)}. \tag{4.7}$$

Once the estimates of the indicator variables (which are not necessarily 0 or 1 any more) are given, the first part of the M step of updating the weights is

evidently

$$\omega_k^{(j+1)} := \frac{1}{N}\sum_{s,t} z_{k,s,t}^{(j+1)}, \tag{4.8}$$

where N is the total number of forecast cases in the training set.

Further, non-linear equations $\frac{\partial \ell_C}{\partial \alpha_k} = 0$ and $\frac{\partial \ell_C}{\partial \beta_k} = 0,\ k = 1, \ldots, K,$ lead us to update formulae

$$\alpha_k^{(j+1)} := \frac{\sum_{s,t} z_{k,s,t}^{(j+1)} \left\{ \left(x_{k,s,t} - \beta_{k,s,t}^{(j)} f_{k,s,t}\right) + \sigma^{(j)} \frac{\varphi\left(\frac{b-\mu_{k,s,t}^{(j)}}{\sigma^{(j)}}\right) - \varphi\left(\frac{a-\mu_{k,s,t}^{(j)}}{\sigma^{(j)}}\right)}{\Phi\left(\frac{b-\mu_{k,s,t}^{(j)}}{\sigma^{(j)}}\right) - \Phi\left(\frac{a-\mu_{k,s,t}^{(j)}}{\sigma^{(j)}}\right)} \right\}}{\sum_{s,t} z_{k,s,t}^{(j+1)}} \tag{4.9}$$

$$\beta_k^{(j+1)} := \frac{\sum_{s,t} z_{k,s,t}^{(j+1)} f_{k,s,t} \left\{ \left(x_{k,s,t} - \alpha_{k,s,t}^{(j)}\right) + \sigma^{(j)} \frac{\varphi\left(\frac{b-\mu_{k,s,t}^{(j)}}{\sigma^{(j)}}\right) - \varphi\left(\frac{a-\mu_{k,s,t}^{(j)}}{\sigma^{(j)}}\right)}{\Phi\left(\frac{b-\mu_{k,s,t}^{(j)}}{\sigma^{(j)}}\right) - \Phi\left(\frac{a-\mu_{k,s,t}^{(j)}}{\sigma^{(j)}}\right)} \right\}}{\sum_{s,t} z_{k,s,t}^{(j+1)} f_{k,s,t}^2}$$

respectively. However, simply using $\mu_{k,s,t}^{(j+1)} := \alpha_k^{(j+1)} + \beta_k^{(j+1)} f_{k,s,t}$ as the location change results in an unstable parameter estimation process due to numerical problems. Thus, analogous to Baran (2014), where the same problem occurred in case studies of wind speed data, a mean correction is applied of the following form

$$\mu_{k,s,t}^{(j+1)} := \mu_{k,s,t}^{(0)} - \sigma^{(j)} \frac{\varphi\left(\frac{a-\alpha^{(j+1)}-\beta^{(j+1)} f_{k,s,t}}{\sigma^{(j)}}\right) - \varphi\left(\frac{b-\alpha^{(j+1)}-\beta^{(j+1)} f_{k,s,t}}{\sigma^{(j)}}\right)}{\Phi\left(\frac{b-\alpha^{(j+1)}-\beta^{(j+1)} f_{k,s,t}}{\sigma^{(j)}}\right) - \Phi\left(\frac{a-\alpha^{(j+1)}-\beta^{(j+1)} f_{k,s,t}}{\sigma^{(j)}}\right)}, \tag{4.10}$$

which reflects to the difference between the location and mean of a truncated normal distribution, see (4.3). Finally, from $\frac{\partial \ell_C}{\partial \sigma} = 0$ we obtain the last update formula

$$\begin{aligned} \sigma^{2(j+1)} := \frac{1}{N}\sum_{s,t}\sum_{k=1}^{K} z_{k,s,t}^{(j+1)} \Bigg\{ & \left(x_{s,t} - \mu_{k,s,t}^{(j+1)}\right)^2 \\ & + \sigma^{(j)} \frac{\left(b - \mu_{k,s,t}^{(j+1)}\right)\varphi\left(\frac{b-\mu_{k,s,t}^{(j+1)}}{\sigma^{(j)}}\right) - \left(a - \mu_{k,s,t}^{(j+1)}\right)\varphi\left(\frac{a-\mu_{k,s,t}^{(j+1)}}{\sigma^{(j)}}\right)}{\Phi\left(\frac{b-\mu_{k,s,t}^{(j+1)}}{\sigma^{(j)}}\right) - \Phi\left(\frac{a-\mu_{k,s,t}^{(j+1)}}{\sigma^{(j)}}\right)} \Bigg\}. \end{aligned} \tag{4.11}$$

Note that without truncation $(-a = b = \infty)$ the terms (4.9) and (4.11) based on $\sigma^{(j)}$ will disappear, so the location (mean) and scale (standard deviation) will be updated separately, no mean correction is needed, and we will get back the classical EM algorithm for normal mixtures.

As a simple alternative, one can ignore the (4.9) update step for α_k and β_k, simplify the mean (4.10) correction step to

$$\mu_{k,s,t}^{(j+1)} := \mu_{k,s,t}^{(0)} - \sigma^{(j)} \frac{\varphi\Big(\frac{a-\mu_{k,s,t}^{(j)}}{\sigma^{(j)}}\Big) - \varphi\Big(\frac{b-\mu_{k,s,t}^{(j)}}{\sigma^{(j)}}\Big)}{\Phi\Big(\frac{b-\mu_{k,s,t}^{(j)}}{\sigma^{(j)}}\Big) - \Phi\Big(\frac{a-\mu_{k,s,t}^{(j)}}{\sigma^{(j)}}\Big)}, \tag{4.12}$$

and only after the EM algorithm stops, estimate location parameters α_k and β_k from a linear regression of the final value of $\mu_{k,s,t}$ on $f_{k,s,t}$.

Finally, one can also try the classical naive method, where the location parameters α_k and β_k are not updated at all, i.e. $\mu_{k,s,t}^{(j+1)} \equiv \alpha_k^{(0)} + \beta_k^{(0)} f_{k,s,t}$.

4.2.3 Truncated normal EMOS model

As reference model for calibrating Box-Cox transformed forecasts for water level, we use a truncated normal distribution $\mathcal{N}_b^a\big(\mu, \sigma\big)$ in (2.4), where the location μ and scale σ are linked to the ensemble members using the following equations

$$\mu := a_0 + a_1 f_1 + \cdots + a_k f_k \qquad \text{and} \qquad \sigma^2 = b_0 + b_1 S^2 \tag{4.13}$$

where S^2 is the ensemble variance (see Hemri and Klein, 2017).

4.2.4 Verification scores

For a truncated normal distribution the CRPS has a simple closed form (see e.g. the `R` package `scoringRules`; Jordan et al., 2019), whereas for the truncated normal mixture (2.1), similar to the mixture model of Baran and Lerch (2016), the second integral expression in the definition (3.1) should be evaluated numerically. Moreover, in our case study each calibration approach provides a predictive CDF F for the Box-Cox transformed water level $X \in [a, b]$. Thus, the CRPS corresponding to the predictive CDF $G(y) := F\big(h_\lambda(y)\big)$ of the original water level $Y = h_\lambda^{-1}(X) \in \big[h_\lambda^{-1}(a), h_\lambda^{-1}(b)\big]$ and a real value y equals

$$\mathrm{CRPS}\,\big(G, y\big) = \int_{h_\lambda^{-1}(a)}^{y} F^2\big(h_\lambda(u)\big)\mathrm{d}u + \int_{y}^{h_\lambda^{-1}(b)} \Big(1 - F\big(h_\lambda(u)\big)\Big)^2 \mathrm{d}u, \tag{4.14}$$

which integral should again be approximated numerically.

In the case study of Section 4.4 we report the p-values of the DM test. p-values less than 0.05 indicate significant difference in scores at a 5 % level.

4.2.5 Analog-based approaches to choosing training data

Following Hemri and Klein (2017), besides the rolling training date we also consider the series distance method (SD; Ehret and Zehe (2011); Seibert et al. (2016)), the hydrograph matching algorithm (HMA; Ewen (2011)), and dynamic time warping (DTW; Sakoe and Chiba (1978)). To mimic the human hydrologist, SD and HMA have an aim of quantifying the gap between two hydrological time series. The SD does a decent job balancing all peaks and troughs as well as the intermediate rising and dropping limbs. This calculation of similarity takes into account adjustments that are necessary to balance the two time series with regard to both amplitude and timing. Similarly, HMA allows for moderate time changes between series, but connects each component of the first to the others. It is also possible to consider the cumulative length of all connections. DTW measures similarity by finding the minimum amplitude error between two time series.

For these purposes, we first locate training data with a forecast time series that is similar to the observations of the considered verification date (generally these methods are used to compare simulated with observed hydrological time series). For each date of verification and each of DTW, HMA, and SD, we pick the 100 most similar training dates from the ECMWF ENS mean trajectories. The verification and training dates with overlapping forecast trajectories are omitted. To read more about these methods, see Hemri and Klein (2017).

4.3 Data

The BMA and EMOS calibration approaches are tested on ensemble water level forecast (cm) at the Rhine River Kaub gauge (546 km) and the corresponding validation observations. Predictions covering an eight-year period from 1 January 2008 to 31 December 2015 are analyzed, with lead times from 1 to 120 hrs and a time step of one hour. This particular gauge has registered the minimum and maximum water levels to be 35 and 825 cm, respectively. The multimodel water level ensemble in our study is formed by plugging forecast predictions for weather variables produced by different ensemble prediction

systems into the hydrological model HBV-96 (a hydrological model designed and operated by the German Federal Institute of Hydrology BfG) (Lindström et al., 1997), which is used for operational runoff forecasting. We consider the ECMWF high-resolution (HRES) forecast, the 51-member ECMWF forecast (ENS) (Leutbecher and Palmer, 2008; Molteni et al., 1996), the 16-member COSMO LEPS forecast of the consortium for small-scale modeling (Montani et al., 2011), and the 11-member NCEP GEFS forecast of the reforecast version 2 of the global ensemble forecast system of the National Center for Environmental Prediction (Hamill et al., 2013). While these EPSs have different ranges, the considered maximal lead time of 5 days is in the intersection of their time spans. A hydrodynamic model is used to transform the runoff forecasts into water level forecasts for the navigation-relevant gauges, including gauge Kaub. Ensemble forecasts are initialized at 6 UTC. The data set in question is part of the data analyzed in Hemri and Klein (2017), where we refer to for more details.

4.4 Results

We apply EMOS and BMA post-processing models for Box-Cox transformed water levels. Similar to Hemri and Klein (2017), a different Box-Cox parameter λ is applied for each lead time (Figure 4.2). The selected parameter maximizes the in-sample skill of the seasonally fitted EMOS models with regard to the CRPS relative to the raw ensemble. As for training, data from the same season of other years is selected. We obtain one estimate for each lead time by taking the mean of the model's estimates over all training periods. The Box-Cox coefficients are applied for all data at all levels (i.e. the observations too). The upper and lower bounds for the truncated normal distribution in BMA and EMOS models, similar to Hemri and Klein (2017), are the Box-Cox transformed bounds of the interval that goes from half the minimum to double the maximum of the recorded water levels. In our case this means the Box-Cox transform of 17.5 and 1,650 cm, respectively.

There's a natural grouping of of the ensemble members, due to the generation process of the hydrological ensemble forecasts. These groups are the ECMWF HRES, ECMWF ENS, COSMO LEPS and NCEP GEFS ensemble forecasts, containing 1, 51, 16 and 11 members respectively. Therefore, Box-Cox transformed water level forecasts are calibrated using the truncated normal BMA model for these exchangeable ensemble members specified by (2.2) and (4.2) and truncated EMOS given by (4.13) and (2.2.2) with $k = 4, M_1 = 1, M_2 = 51$, $M_3 = 16$ and $M_4 = 11$. This results in 12 parameters to estimate for the

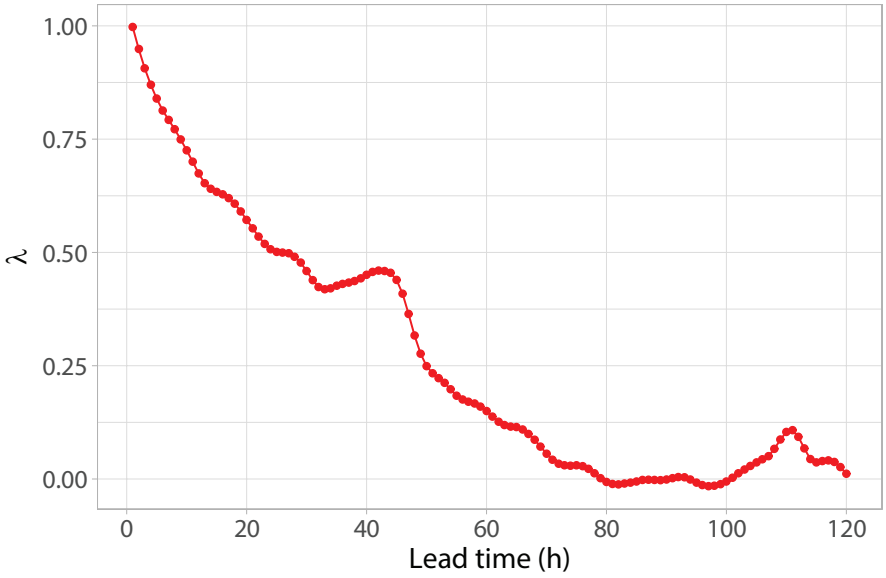

Figure 4.2: Box-Cox transformation parameter λ as function of the lead time.

BMA model, whereas the corresponding EMOS model has 7 parameters only. To ensure stability of parameters estimation, both BMA and EMOS models are trained using forecasts and observations of 100 days. We consider RTPs and analog-based approaches.

The quantile-based scores (MAE, coverage and average width) are evaluated using the inverse Box-Cox transformation of the observations and forecasts. This ensures comparability, since the BMA and EMOS models are fit to Box-Cox transformed values.

4.4.1 BMA Vs. EMOS using RTP

To check the forecast skill of the truncated normal BMA model introduced in section 4.2.1, first the standard approach in BMA and EMOS post-processing (Gneiting et al., 2005; Raftery et al., 2005) is applied, where model parameters are estimated using RTP of 100 days. This means that BMA and EMOS models are verified for the period from 10 April 2008 to 31 December 2015 (2822 calen-

dar days). We consider 1 day ahead calibration for all lead times, which means, that for example, to model the water level for the 1 of January 2015, we use forecasts and observations for the preceding 100 days ending at 31 December 2014. For 24-hour and 120-hour lead times, the last forecasts are initialized at 30 December 2014 and 26 December 2014 respectively.

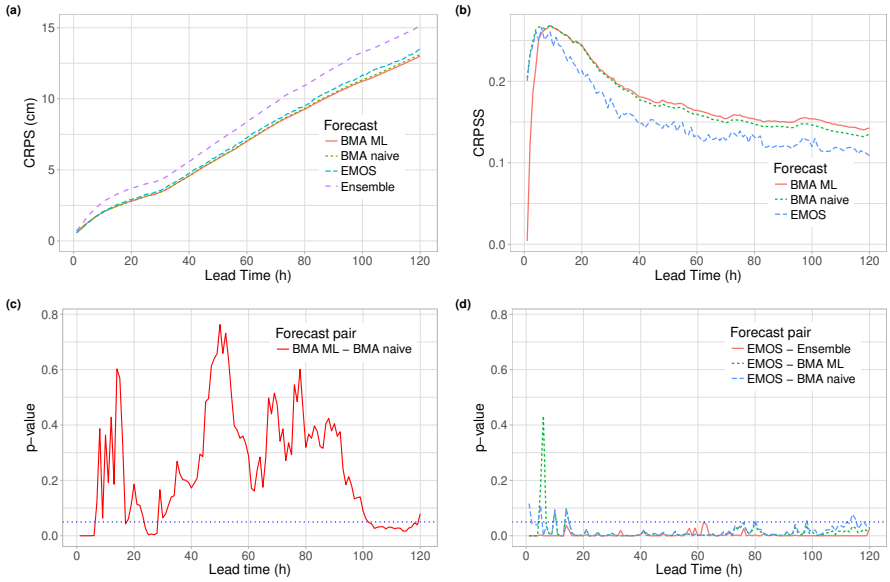

Figure 4.3: Mean CRPS values (a) and CRPSS with respect to the raw ensemble (b); p-values of DM tests for equality of mean CRPS of the two BMA approaches (c) and of all models compared to EMOS (d). Horizontal dotted lines of (c) and (d) indicate a 5 % level of significance.

In Figure 4.3a, the mean CRPS values of all post-processing approaches with the raw ensemble are plotted as functions of the lead time. Compared to the raw ensemble, all calibration approaches reduce the mean CRPS and the gap increases with the lead time. The difference between the forecast skills is more clear in Figure 4.3b that shows the CRPSS values with respect to the raw ensemble forecast. All presented methods have their maximal skill score at hour 9. This means that the relative gap in CRPS between raw and post-processed forecasts increases up to hour 9 and decreases again thereafter. This does not

mean that the absolute forecast skill increases with lead time between hours 1 and 9. For short lead times, this increase is very fast and naive BMA shows the best predictive performance, whereas for longer lead times, the pure ML BMA start outperforming the competitors. When comparing predictive performance, we should take into account the obvious association of longer lead times with larger forecast uncertainty. The DM tests for equal predictive performance show that naive BMA significantly outperforms the ensemble for all lead times, whereas for pure ML BMA the same holds except for hour 1. In terms of mean CRPS, the two BMA approaches differ significantly for very short and long lead times. This is reflected on the graph of p values displayed in Figure 4.3c. As depicted in Figure 4.3d, EMOS also outperforms significantly the raw ensemble for all lead times except for the first few hours where it underperforms the BMA approaches.

In terms of the MAE, there's much less variety in the performance of BMA and EMOS calibrated medians. Showing the different MAE values with respect to the raw ensemble, Figure 4.4a tells us that the pure ML BMA has the best forecast skill, but it underperforms the raw ensemble until hour 70. DM tests for equality of MAE values tell that all differences shown from the raw ensemble in Figure 4.4a are significant (DM test results are not shown), while it's not the case if a comparison of the performance of the three post-processing methods is made. To see the p-values, please check Figure 4.4b.

In Figure 4.4c, the coverage of nominal 97.5% central prediction intervals as function of the lead time are shown. Clearly, we can see the beneficial effect of post-processing on calibration. For all lead times, all post-processing approaches result in an almost perfect coverage, while the coverage of the raw ensemble is much lower and heavily dependent on the lead time. In terms of coverage the two BMA approaches are almost identical, and after hour 4 they are more closer to the nominal value than those of EMOS after hour 4. Finally, as shown in Figure 4.4d, for all lead times, the raw ensemble produces the sharpest forecasts, however, this is at the cost of being uncalibrated.

This is fully in accordance with the verification rank histograms of the raw ensemble and PIT histograms of post-processed forecasts, as shown in Figure 4.5a for lead times 24, 72 and 120 hours. All verification rank histograms for all lead times are U-shaped, which means that the raw ensemble is strongly underdispersive and requires post-processing. The statistical calibration of the forecast is greatly improved by using BMA and EMOS approaches. This is reflected more on the uniform shape of the PIT histograms, whilst the naive BMA and EMOS still shows a small underdispersion at 120 hour lead time. Kolmogorov-Smirnov's test statistic values for the uniformity of the PIT values

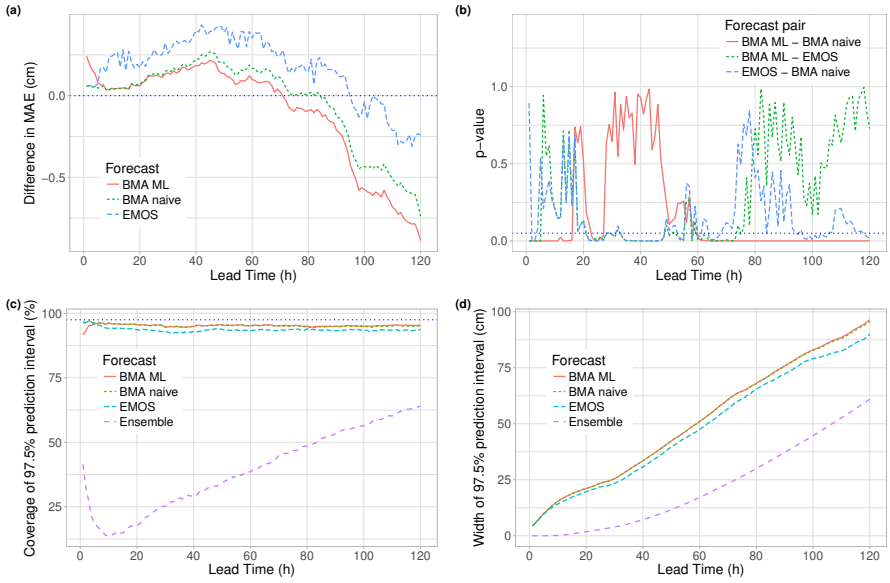

Figure 4.4: Top row: Difference in MAE values from the raw ensemble (a) and p-values of DM tests for equality of MAE of the various post-processing approaches (b). Horizontal dotted lines indicate the reference raw ensemble (a) and a 5 % level of significance (b). Bottom row: Coverage (c) and average width (d) of nominal 97.5 % central prediction intervals. In panel (c) the ideal coverage is indicated by the horizontal dotted line.

for all post-processing approaches are presented in Figure 4.5b (the better the fit, the smaller the test statistic). Figure 4.5b nicely demonstrates the ranking of all approaches in terms of goodness of fit, even though the uniformity of the PIT values of naive BMA, ML BMA and EMOS can be accepted at a 5% level of significance for only 6 (4,5,6,7,14 and 17 hrs), 9 (5,6,7,14,17,72,75,77 and 79 hrs) and 4(5,6,7 and 9 hrs) different lead times, respectively.

4.4.2 Analog-Based Vs. RTP

In comparison to the use of RTP methods of selecting training data, Hemri and Klein (2017) found that the analog-based ones undoubtedly improve the predictive performance of EMOS. Here we examine whether this holds for doubly

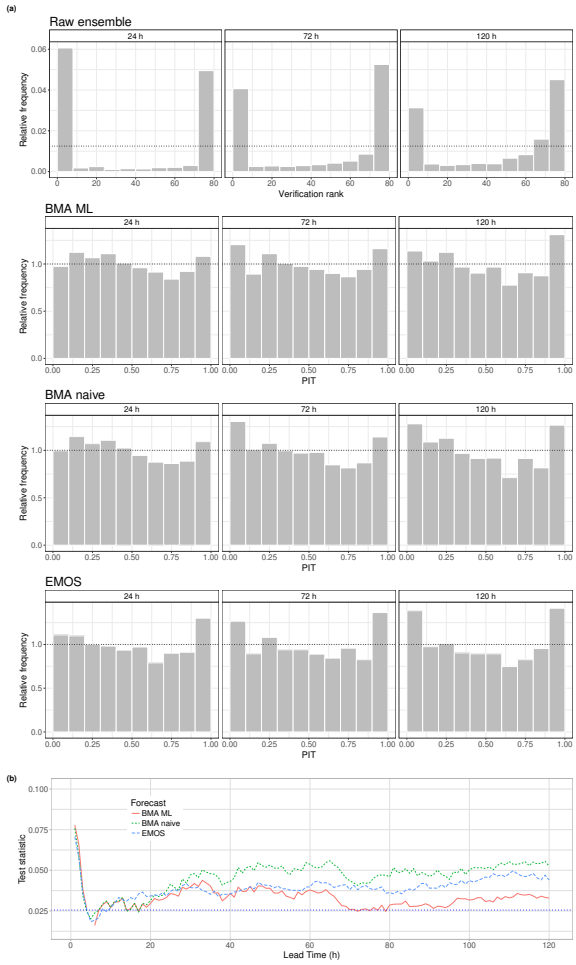

Figure 4.5: Verification rank histogram of the raw ensemble and PIT histograms of the BMA and EMOS post-processed forecasts for lead times 24, 72 and 120 hours (a); values of the test statistic of Kolmogorov-Smirnov tests for uniformity of PIT values (b). Smaller values indicate better fit, dotted horizontal line corresponds to 5 % level of significance.

truncated normal BMA or not. We analyze the forecast skill of the pure ML BMA with training data selected according to the DTW, HMA and SD approaches described in section 4.2.5 (BMA DTW, BMA HMA and BMA SD) and using BMA RTP. To guarantee equivalence, all models are verified on the same data (2,822 calendar days, from 10 April 2008 to 31 December 2015 as in section 4.4.1).

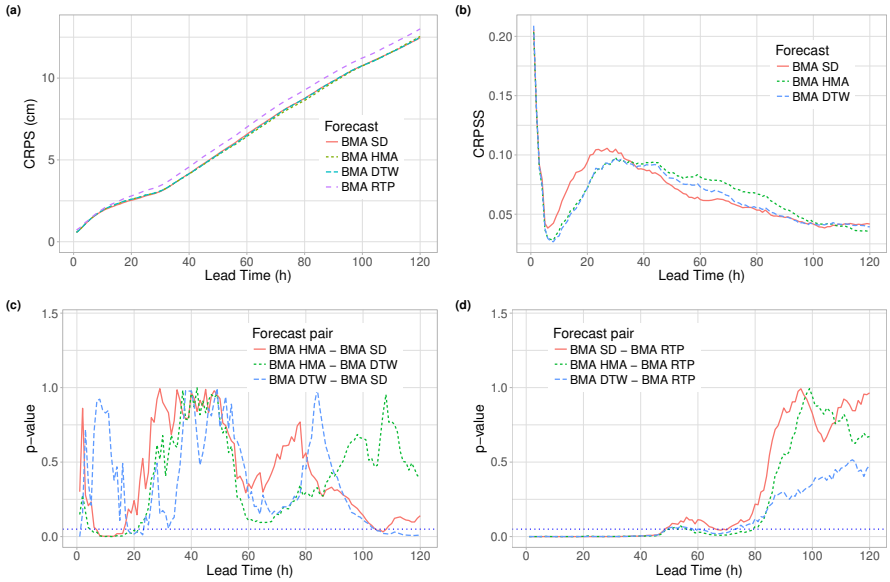

Figure 4.6: Mean CRPS values of BMA forecasts (a) and CRPSS with respect to the BMA with rolling training period (BMA RTP) (b); p-values of DM tests for equality of mean CRPS of the analog-based BMA approaches (c) and analog-based models compared to BMA RTP (d). Horizontal dotted lines of (c) and (d) indicate a 5 % level of significance.

The mean CRPS values of the four methods in question as function of the lead time, are plotted in Figure 4.6a, while the analogous skill scores with respect to the BMA RTP are shown in Figure 4.6b. The DM tests of equal predictive performance in terms of the mean CRPS are reported in Figures 4.6c and 4.6d. They show that, after hour 50 all of the analog-based methods outperform the BMA RTP significantly and there's no big difference between the various

analog-based approaches.

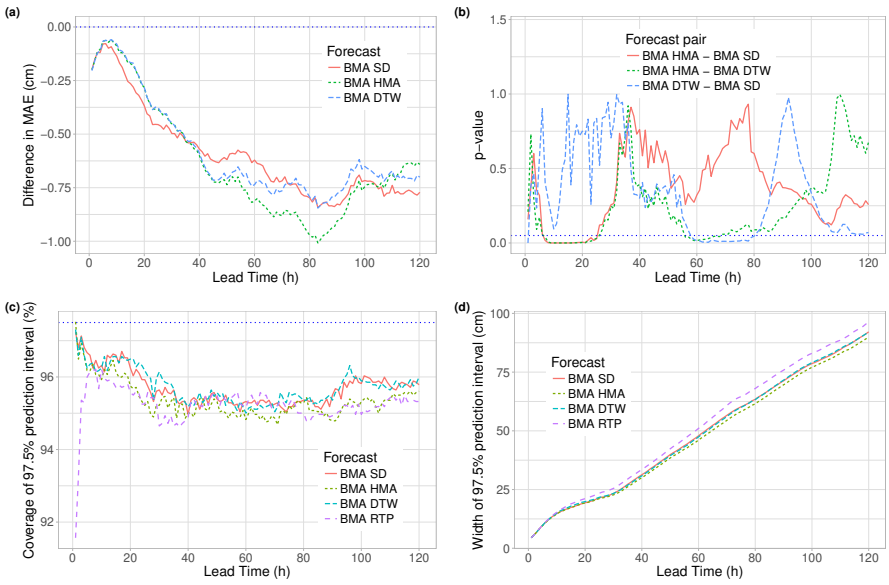

Figure 4.7: Top row: Difference in MAE values from the BMA RTP (a) and p-values of DM tests for equality of MAE of the various analog-based BMA approaches (b). Horizontal dotted lines indicate the reference BMA RTP (a) and a 5 % level of significance (b). Bottom row: Coverage (c) and average width (d) of nominal 97.5 % central prediction intervals. In panel (c) the ideal coverage is indicated by the horizontal dotted line.

Figure 4.7a shows that the analog-based training methods have more gain in terms of MAE. For all lead times, the MAE values of the analog-based methods are significantly lower thant those of the BMA RTP (DM test results are not reported). For lead times until 35 hr the BMA SD, for lead times between 40 and 104 hrs the BMA HMA, whereas for longer lead times again the BMA SD yields the smallest MAE values. But the difference in MAE is not significant after 80 hours for all analog-based methods, as shown in Figure 4.7b.

The coverage values reported in Figure 4.7c indicate that all analog-based models outperform the BMA RTP for very short lead times, while there is no big difference within the analog-based models, moreover, BMA DTW, BMA

HMA and BMA SD models have a slightly sharper central prediction intervals than BMA RTP (Figure 4.7d).

Finally, as shown in Figure 4.8a, the PIT histograms of the analog-based models are similar to the corresponding histograms in the second row of Figure 4.5a. They indicate similar calibration of all BMA approaches for lead times 24, 72 and 120 hrs. The behaviour of PIT values can be more investigated by comparing Figure 4.8b, where the values of the test statistic of the Kolmogorov-Smirnov test for uniformity for the analog-based BMA approaches are plotted, with the corresponding line (BMA ML) of Figure 4.8b. The uniformity of PIT of BMA SD, BMA HMA, and BMA DTWcan be accepted at a 5% level of significance for 25, 21, and 24 lead times, respectively, whereas for the BMA RTP (BMA ML in Figure 4.5b) there are just 9 such lead times.

As detailed previously, BMA RTP is significantly outperformed by the analog-based BMA approaches at shorter lead times up to about 40 hour. Up to around 80 hours this outperformance is still borderline significant, while this isn't the situation for longer lead times. This means that only for shorter lead times we have benefit by the analog-based models. This is not surprising so far as that a higher predictability at shorter lead times, which is usually the case compared to longer lead times, implies also that similarity in forecasts is more likely to be connected to similarity in the observations. Any analog-based approach assumes that such a connection can be established.

4.4.3 Analog-Based BMA Vs. Analog-Based EMOS

The results of section 4.4.1 state that, when the coefficients estimation of the post-processing model is based on RTP, pure ML BMA significantly outperforms EMOS for almost all lead times (Figures 4.3b and 4.3d) in terms of the mean CRPS and MAE for very short (1-5 hr) and medium (21-75 hr) lead times (Figures 4.4a and 4.4b).

This situation changes significantly when using analog-based training data. Figure 4.9a shows the CRPSS values of the analog-based BMA models with respect to the corresponding EMOS ones, as a function of the lead time. The skill scores range between -0.0117 and 0.0084 (very short interval), and BMA DTW , BMA HMA and BMA SD outperform their EMOS counterpart for only 80, 82 and 60 different lead times respectively; however, for lead times 27, 28, 31, 43–46, and 71–117 hr none of the differences are significant (see Figure 4.9b). In this way, for SD, HMA, and DTW approaches there are 46, 47, and 41 different lead times, respectively, when BMA is significantly better in terms of mean CRPS and 11, 8, and 14 when EMOS performs better.

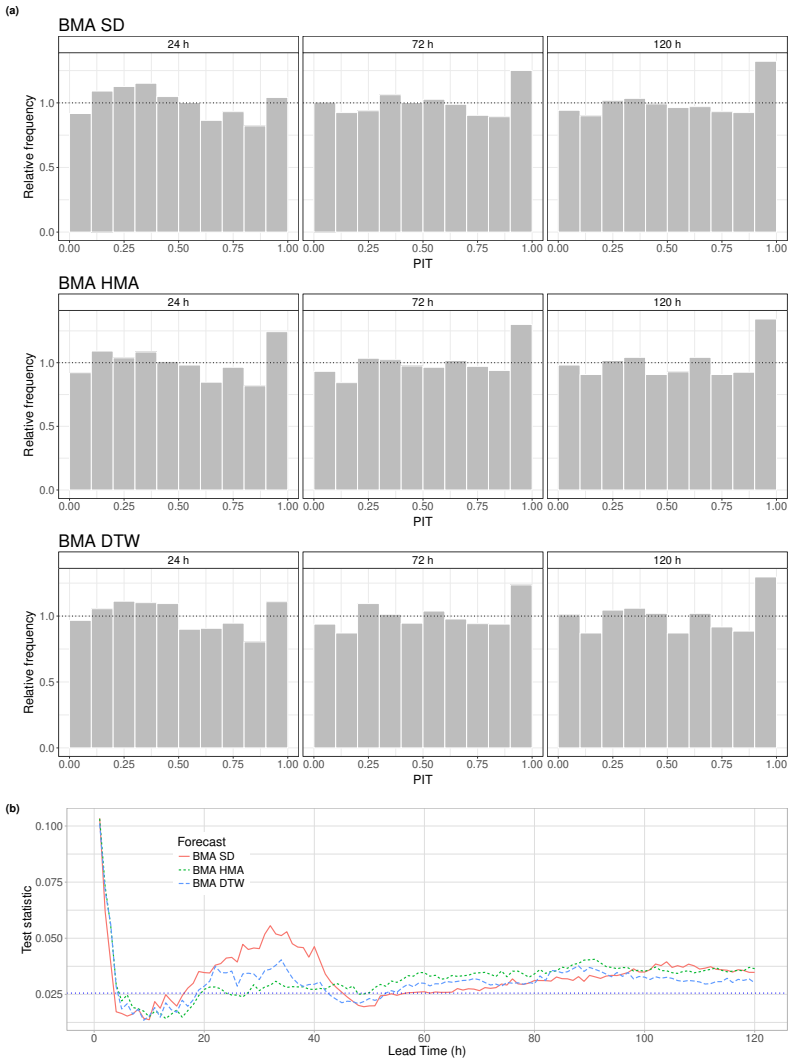

Figure 4.8: PIT histograms of the analog-based BMA post-processed forecasts for lead times 24, 72 and 120 hours (a); values of the test statistic of Kolmogorov-Smirnov tests for uniformity of PIT values (b). Smaller values indicate better fit, dotted horizontal line corresponds to 5 % level of significance.

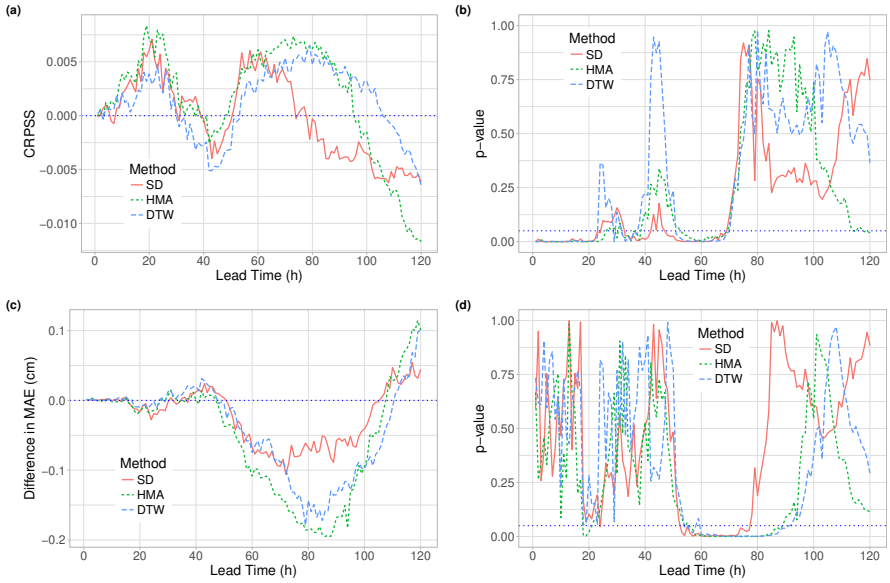

Figure 4.9: Top row: CRPSS values of the analog-based BMA models with respect to the corresponding analog-based EMOS approaches (a); p-values of DM tests for equality of mean CRPS of BMA and EMOS post-processed forecasts (b). Horizontal dotted line of panel (b) indicates a 5 % level of significance. Bottom row: Difference in MAE values of various analog-based BMA models from the corresponding EMOS approaches (c) and p-values of DM tests for equality of MAE (d). Horizontal dotted lines indicate the reference EMOS model (c) and a 5 % level of significance (d).

BMA DTW, BMA HMA and BMA SD have lower MAE values than EMOS DTW, EMOS HMA and EMOS SD in 77, 82 and 77 cases respectively (see Figure 4.9c). Moreover, at a 5% significance level, as indicated in Figure 4.9d, the corresponding DM tests show no big difference at most of the cases.

Unlike the case of RTPs, the differences are more less clear in this situation, but the analog-based BMA approaches still perform slightly better than their EMOS counterparts. Since EMOS is based on a simple parametric predictive distribution, it's clear why it gains a lot from the analog-based selection of training periods. This is most likely because of the fact that analog-based training periods allow for a strong dependence between raw ensemble means and the corresponding observations. As Hemri and Klein (2017) stated, this improves sharpness, at the cost of slightly deteriorating calibration. In contrast, BMA ends in a pretty flexible predictive distribution, which in our case study can have several modes. This flexibility results in a robust improvement in forecast skill in comparison to EMOS when RTP selection of training periods is used. Using analog-based selection of training periods sharpens the BMA predictive distribution; however, its impact on forecast skill is lighter than in the case of EMOS. This may also imply that through the usage of analog-based training periods, the post-processed forecast skill is near the theoretically achievable skill based on the available forecast model used here, and accordingly, the benefit through the usage of the more sophisticated BMA is rather marginal.

4.5 Conclusion

Post-processing often increases the calibration of probabilistic and accuracy of point forecasts when compared to the raw ensemble. The BMA model outperforms the reference EMOS method significantly with rolling window training data. When using analog-based training period selection, the difference in forecast skill narrows dramatically, leaving BMA with only a minor advantage over EMOS. With analog-based training period, we suggest the use of BMA model when the data set at hand is short and EMOS model otherwise.

Further, following the ideas of Hemri et al. (2015) and Bellier et al. (2018), one can combine the BMA calibrated forecasts corresponding to different locations and lead times either into temporally, or both spatially and temporally coherent multivariate predictions with the help of modern calibration techniques such as e.g. the ensemble copula coupling (Schefzik et al., 2013) or the Gaussian copula approach (Pinson and Girard, 2012).

Chapter 5

Post-processing of solar irradiance

5.1 Introduction

Transitioning to renewable energy sources such as wind and solar power is critical for reducing greenhouse gas emissions (Van der Meer et al., 2018). For incorporating dynamic power systems into the electrical grid in order to balance demand and supply, accurate and reliable forecasts of power generation from those sources are becoming increasingly relevant (Gottwalt et al., 2016; González Ordiano et al., 2020).

There are four major factors that limit solar irradiation (Bozonnat and Schlosser, 2014). First the geometry, revolution, and rotation of the earth. Second the gases, liquids, and solid particles in the atmosphere. The third factor is cloud attenuation. While the fourth factors are the terrain elevation, shadow, surface orientation, and inclination. In this chapter, we consider 7 and 3 representative observation locations for Hungary and Germany respectively. In Hungary, the highest recorded radiation is in July, while the lowest in December. This is because the amount of cloud cover is less in July (even though daylight does not last as long as in June). Moreover, the maximum cloud cover and short daylight is in December. Typically, southern regions receive more solar irradiance (of 1334–1362 kWh/m^2 as average annual global radiation) than northern regions (less than 1195 kWh/m^2 as average annual global radiation) (`www.met.hu`). For Germany, the highest recorded global ir-

radiance is in June (169-224 kWh/m^2) while the lowest is in December (12-34 kWh/m^2) (`www.dwd.de` as monthly global radiation).

Using corresponding NWP ensemble predictions as data, we propose post-processing models for various measured forms of solar irradiance, based on the EMOS approach. We use a censored logistic forecast distribution inspired by similar models for post-processing ensemble forecasts of precipitation accumulation (Scheuerer, 2013; Baran and Nemoda, 2016), to account for the specific discrete-continuous aspect of solar irradiance due to the positive probability of observing zero irradiance during night-time. For lead times of up to 48 and 120 hrs, respectively, the post-processing models are used in two case studies that concentrate on different solar irradiance variables, NWP models, temporal resolutions, and geographic regions (Hungary and Germany). We also look at various temporal compositions of training datasets for model estimation and use periodic models to better capture seasonal variance in solar irradiance.

The remainder of this chapter, which is based on Schulz et al. (2021), is organized as follows. We start with the proposed post-processing approach in Section 5.2 followed by an introduction of the datasets at hand in Section 5.3. Finally, results for the two case studies is presented in Section 5.4 and the chapter ends with some conclusions. The results of this chapter have been published in Baran et al. (2021).

5.2 Post-processing methods

A distribution-based approach, in general, is less capable of representing complex forecast distributions than a non-parametric one, but it is computationally more efficient and particularly well suited in sparse data settings. We chose the distribution-based EMOS approach due to the limited amount of data. In the following section we propose a new model based on a censored logistic distribution.

5.2.1 Choice of forecast distribution

Because of the discrete-continuous nature of solar irradiance, non-negative predictive distributions with positive mass for zero irradiance are needed. Similar to parametric approaches to post-processing ensemble forecasts of precipitation accumulation, one can either choose the more complex method of mixing a point mass at zero and a suitable continuous distribution with non-negative support (Sloughter et al., 2007; Bentzien and Friederichs, 2012), or left-censor an ap-

propriate continuous distribution at zero (see e.g. Scheuerer, 2013; Baran and Nemoda, 2016). In this chapter, we'll focus on the latter and implement the censored logistic distribution that our approach is based on.

Consider a logistic distribution $\mathcal{L}(\mu, \sigma)$ with location μ and scale $\sigma > 0$ specified by the PDF

$$g(x; \mu, \sigma) := \frac{\mathrm{e}^{-(x-\mu)/\sigma}}{\sigma \left(1 + \mathrm{e}^{-(x-\mu)/\sigma}\right)^2}, \qquad x \in \mathbb{R},$$

and the CDF $G(x; \mu, \sigma) := \left(1 + \mathrm{e}^{-(x-\mu)/\sigma}\right)^{-1}$.

The logistic distribution left-censored at zero (CL0) assigns point mass $G(0; \mu, \sigma) = \left(1 + \mathrm{e}^{\mu/\sigma}\right)^{-1}$ to the origin, i.e. the probability of observing a negative value (before censoring) is the probability of observing zero afterwards. The CL0-distribution can be defined by the CDF

$$G_0^c(x; \mu, \sigma) := \begin{cases} G(x; \mu, \sigma), & x \geq 0, \\ 0, & x < 0. \end{cases} \tag{5.1}$$

or the generalized PDF

$$g_0^c(x; \mu, \sigma) = \mathbb{1}_{\{x=0\}} G(0; \mu, \sigma) + \mathbb{1}_{\{x>0\}} g(x; \mu, \sigma). \tag{5.2}$$

The p-quantile q_p $(0 < p < 1)$ of (5.1) equals 0 if $p \leq G(0; \mu, \sigma)$ and $q_p = \mu - \sigma\big(\log(1-p) - \log p\big)$, otherwise. With the help of (5.2) it is straightforward to show that the corresponding mean equals

$$\mu_0^c := \mu + \sigma \log\left(1 + \mathrm{e}^{-\mu/\sigma}\right).$$

Initial tests on the ICON-EPS dataset detailed in Section 5.3.2 were conducted with a censored normal predictive distribution; however, the results indicated that the proposed CL0-EMOS method results in a slightly improved predictive performance. In post-processing research, the choice of parametric families for the forecast distribution has been critical, see for example, Yang (2020b), Yagli et al. (2020), and Le Gal La Salle et al. (2020) for considerations in the sense of solar irradiance forecasting. More involved approaches based on mixtures or combinations of several forecast distributions proposed in the meteorological literature (Baran and Lerch, 2016, 2018) may be able to improve performance, but they increase model complexity and computational costs.

5.2.2 Ensemble model output statistics models for solar irradiance forecasting

As before, for a given location, time point, and forecast horizon, denote the ensemble member forecasts of solar irradiance by $f_1, f_2, \ldots, f_K$. In the simplest proposed EMOS model, the location parameter μ and the scale parameter σ of the CL0-distribution are connected to the ensemble members via the following link functions

$$\mu = a_0 + a_1 f_1 + \cdots + a_K f_K + \nu p_0 \qquad \text{and} \qquad \sigma = \exp\big(b_0 + b_1 \log S^2\big), \quad (5.3)$$

where p_0 and S^2 are the proportion of zero observations and the ensemble variance, respectively, that is

$$p_0 := \frac{1}{K}\sum_{k=1}^{K} \mathbb{1}_{\{f_k=0\}} \qquad \text{and} \qquad S^2 := \frac{1}{K-1}\sum_{k=1}^{K}(f_k - \overline{f})^2.$$

As before, EMOS coefficients $a_0, a_1, \ldots, a_K, \nu$ and b_0, b_1 are estimated according to the optimum score principle of Gneiting and Raftery (2007), i.e., by optimizing the mean value of an appropriate verification metric (usually the CRPS defined by (3.1)) over the training data consisting of past pairs of forecasts and observations for a given time period.

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

$$y_t = c_0 + c_1 \sin\left(\frac{2\pi t}{365}\right) + c_2 \cos\left(\frac{2\pi t}{365}\right) + \varepsilon_t \qquad \text{and} \quad (5.4)$$

$$y_t = d_0 + d_1 \sin\left(\frac{2\pi t}{365}\right) + d_2 \cos\left(\frac{2\pi t}{365}\right) + d_3 \sin\left(\frac{4\pi t}{365}\right) + d_4 \cos\left(\frac{4\pi t}{365}\right) + \varepsilon_t, \quad (5.5)$$

where 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 duration of length n. One can calculate the h ahead predictions $\widehat{y}$ and $\widehat{f_k}$ of the observation and ensemble members, respectively, using either (5.4) or (5.5), and consider the following modified link function for the location:

$$\mu = \widehat{y} + a_0 + a_1\big(f_1 - \widehat{f_1}\big) + \cdots + a_K\big(f_K - \widehat{f_K}\big) + \nu p_0. \quad (5.6)$$

Under the assumption that each ensemble member can be identified and tracked, the model formulations (5.3) and (5.6) are valid. However, both ICON-EPS and AROME-EPS discussed in details in Section 5.3 have exchangeable groups of ensemble members. The exchangeable versions of link functions (5.3) and (5.6) are

$$\mu = a_0 + a_1\overline{f}_1 + \cdots + a_K\overline{f}_K + \nu p_0 \tag{5.7}$$

and

$$\mu = \widehat{y} + a_0 + a_1\big(\overline{f}_1 - \widetilde{f}_1\big) + \cdots + a_K\big(\overline{f}_K - \widetilde{f}_K\big) + \nu p_0, \tag{5.8}$$

respectively, where $\widetilde{f}_k$ is the prediction of $\overline{f}_k$ for lead time h based either on (5.4) or (5.5).

5.3 Data

In the case studies of Section 5.4, we test the EMOS models proposed in Section 5.2.2 using forecasts of different types of solar irradiance provided by the following two different EPSs, which cover distinct forecast domains. The locations of observation stations used in this case study are shown in Figure 5.1 for both AROME-EPS and ICON-EPS.

5.3.1 AROME-EPS

The HMS operates the 11-member Applications of Research to Operations at Mesoscale EPS (AROME-EPS) which spans the Transcarpatian Basin with a horizontal resolution of 2.5 km (Jávorné Radnóczi et al., 2020). It is made up of ten exchangeable ensemble members derived from perturbed initial conditions, as well as a control member derived from an unperturbed analysis. For seven representative locations in Hungary (Aszód, Budapest, Debrecen, Kecskemét, Pécs, Szeged, Tápiószele), the dataset at hand includes ensemble forecasts of instantaneous values of global horizontal irradiance (GHI) (W/m^2) for grid points closest to these stations along with the corresponding validation observations of the HMS for the period between 7 May 2020 and 14 October 2020. Forecasts are initialized at 00 UTC and have a prediction horizon of up to 48 hrs with a temporal resolution of 15 minutes. However, as observations are available in every 10 minutes, in the following analysis, a 30 minutes common time step is applied.

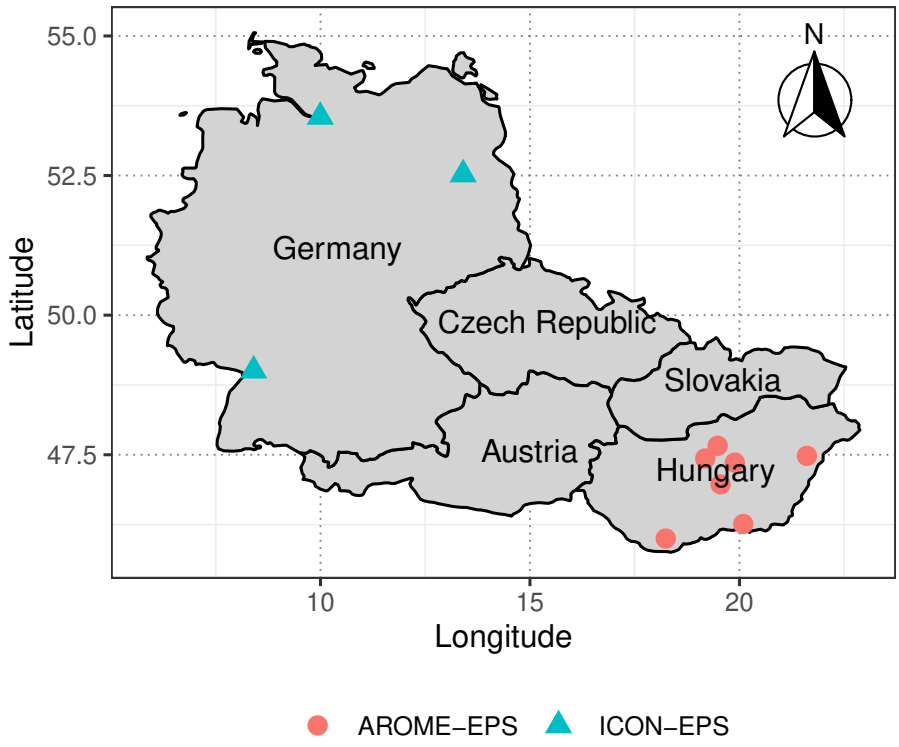

Figure 5.1: Location of observation stations on the map for both AROME-EPS and ICON-EPS.

5.3.2 ICON-EPS

The 40-member global ICOsahedralNonhydrostatic EPS (ICON-EPS; Zängl et al., 2015) of the German Meteorological Service (DWD; Deutsche Wetterdienst) was launched in 2018 and has a horizontal resolution of 20 km over Europe (ICON-EU EPS). The ensemble members are generated using perturbations in initial conditions and in various physics tuning parameters, and the forecasts are updated four times per day at 00/06/12/18 UTC with lead times of up to 120 hrs (Reinert et al., 2021). Hourly forecasts are available up to 48 hrs, while 3 hr forecasts for lead times 51 to 72 hrs, and 6 hr forecasts for lead times

78 to 120 hrs. Ensemble forecasts of the two components of GHI: direct normal irradiance (DNI) calibrated for the solar zenith angle θ (i.e., $\text{DNI} \cdot \cos(\theta)$) and diffuse horizontal radiation (DHI) (W/m^2) are included in our dataset. From the Open Data Server of DWD (DWD Climate Data Center, 2020), we also obtained corresponding observational data for weather stations located near the major cities of Berlin, Hamburg, and Karlsruhe. ICON presently offers three ways for horizontal interpolation of data from a native triangular grid to a latitude-longitude grid (Reinert et al., 2021): radial basis functions, barycentric interpolation and nearest-neighbor interpolation. The observations are computed based on 10-minute sums of the corresponding variables. For more details we refer to Becker and Behrens (2012) for the observations and Reinert et al. (2021) for the ensemble predictions.

The ICON ensemble forecasts are given as averages over the time interval from the previous lead time to the lead time of interest, for example, the forecast for a 12 hr lead time is the average predicted irrandiance between 11 and 12 hr after initialization time, and the forecast for a 90 hr lead time is the average between 84 and 90 hr. The entire dataset used here covers the period 27 December 2018 – 31 December 2020.

In the following, we will refer to GHI as global irradiance, $\text{DNI} \cdot \cos(\theta)$ as direct irradiance, and DHI as diffuse irradiance to simplify the distinction between the different forms of irradiance.

5.4 Results

Please note that ensemble predictions are available for both datasets over several lead times, as well as four separate initialization times of the NWP model in the case of the ICON-EPS dataset. When estimating model parameters, these are handled separately, i.e., a separate post-processing model is calculated for each lead and initialization period, based on training datasets that only contain data from such lead and initialization times. By ensuring that the training data covers the same time of day as the observation, we aim to account for changes in the forecast error characteristics of the raw ensemble predictions over time, as well as possible diurnal effects. When using (5.4) or (5.5), seasonal variations for a given time of day, such as the effects of different solar zenith angles, are implicitly modeled.

The forecast skill of various variants of the CL0-EMOS model introduced in Section 5.2.2 is evaluated in the following case studies. For the HMS AROME-EPS ensemble forecasts of GHI, we first consider a simple CL0-EMOS variant,

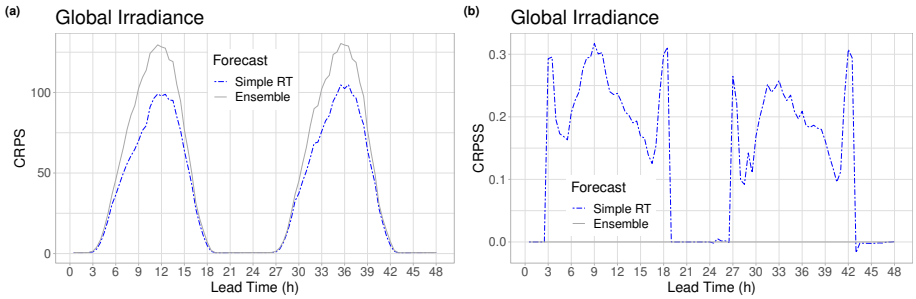

Figure 5.2: Mean CRPS of regionally post-processed and raw ensemble forecasts of GHI (a) and CRPSS with respect to the raw ensemble (b) as functions of lead time for the AROME-EPS dataset.

and then we investigate the performance of the more complicated models for the DWD ICON-EPS ensemble forecasts of direct and diffuse irradiance.

5.4.1 Results for the AROME-EPS dataset

As discussed in Section 5.3.1, the AROME-EPS consists of a control member and 10 exchangeable ensemble members obtained using perturbed initial conditions. In comparison to the ICON-EPS dataset particularly, the dataset at hand only spans a short time period, with forecast-observation pairs available for only 159 calendar days.

As a result, the available training periods are insufficient for accurate modeling of seasonal oscillations, we only consider a CL0-EMOS model in which the ensemble members are connected to the location through (5.7) with $K = 2$ and $M_1 = 1$, $M_2 = 10$, requiring the estimation of six parameters. We consider regional estimation with a 31-day RTP and 127 calendar days for forecast verification (9 June 2020 – 13 October 2020), and label this model the *simple RT* model. The choice of the training period length corresponds to typical values in the post-processing literature and was made to have a similar forecast case per parameter ratio as for the best performing model of Section 5.4.2. Given the dataset's small scale, it's not surprising that using monthly expanding training periods or local parameter estimation procedures leads to poor predictive performance; hence, we've omitted the corresponding results.

Remember that the global irradiance ensemble predictions are given with a 30-minute temporal resolution. Since all AROME-EPS forecasts initialized at

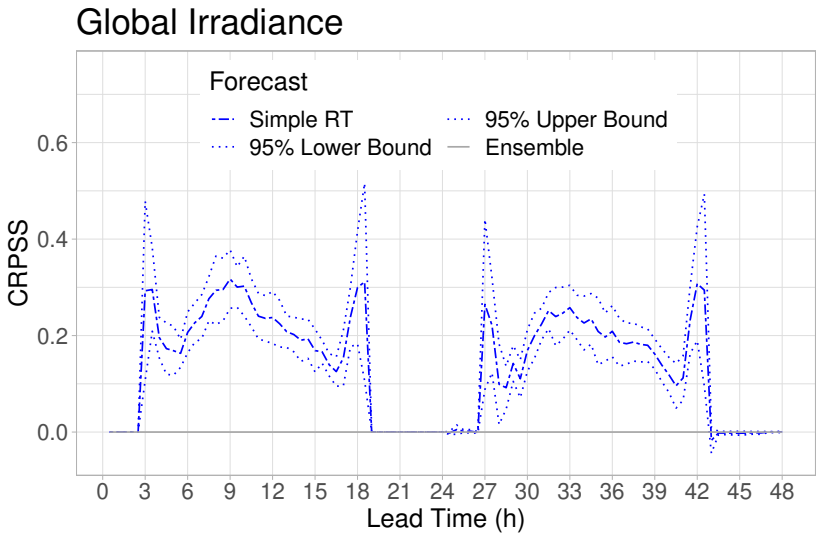

Figure 5.3: CRPSS of EMOS regional post-processed forecasts with respect to the raw ensemble together with 95 % confidence intervals for the AROME-EPS dataset.

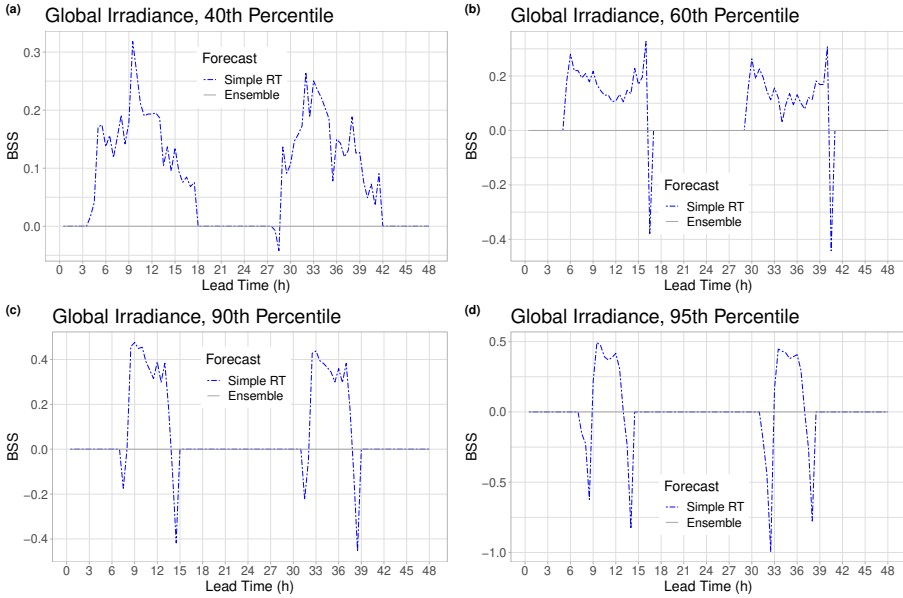

Figure 5.4: BSS of regionally post-processed forecasts with respect to the raw ensemble as function of lead time for the AROME-EPS dataset.

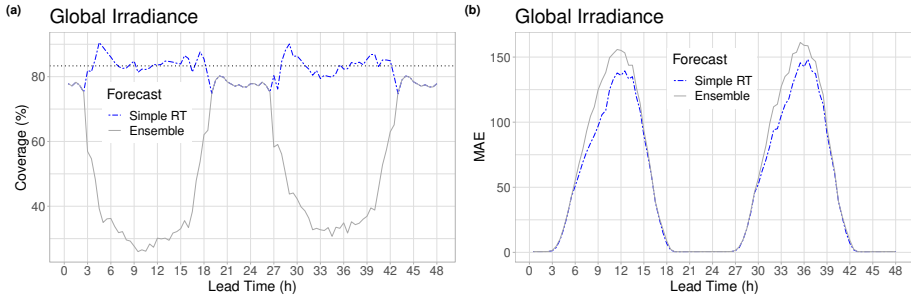

Figure 5.5: Coverage of the nominal 83.33 % central prediction intervals of regionally post-processed and raw forecasts (a); MAE of the median forecasts (b) for the AROME-EPS dataset.

00 UTC, the forecast lead time is either the same as the observation time or has a 24-hour shift. As a result, all scores are expressed as functions of lead time. We then take the average of the results from all seven observation locations over Hungary.

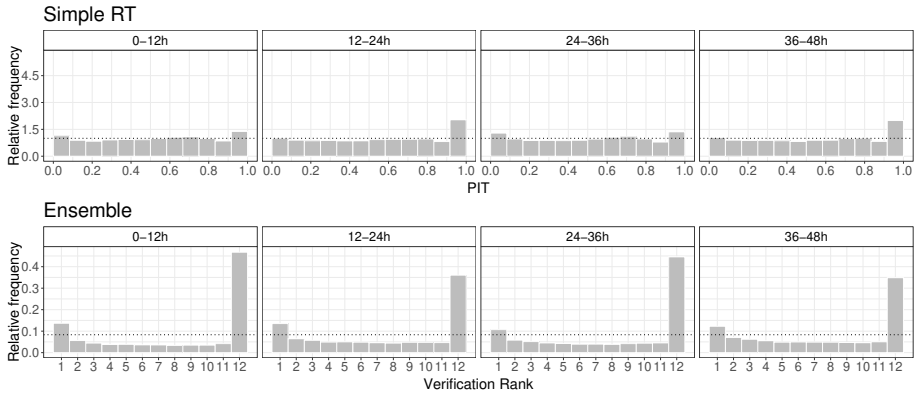

Figure 5.6: PIT histograms of regionally post-processed and verification rank histograms of raw ensemble forecasts of global irradiance for lead times 1-12h, 12-24h, 24-36h and 36-48h.

The mean CRPS of post-processed and raw ensemble forecasts, as well as the CRPSS with respect to the raw ensemble, are shown in Figure 5.2. When positive irradiance is likely to be observed (3 – 19 UTC), post-processing using the simple RT method enhances forecast performance; otherwise, no corrections are made, resulting in a skill score of zero. We used 2 000 block bootstrap samples (resampling) using the stationary bootstrap scheme with mean block length determined according to Politis and Romano (1994), to compute 95% percent confidence intervals to determine the statistical significance of the improvements in predictive performance relative to the raw ensemble predictions, and found that the observed improvements are statistically significant (Figure 5.3). The large jumps in the CRPSS at 4, 19, 27 and 42 hr are mainly caused by numerical issues, as at these lead times the mean CRPS of both raw and post-processed forecasts is very close to 0, and also leads to an increased width of the confidence intervals (Figure 5.3). One can acquire qualitatively similar observations in a related context e.g. in Bakker et al. (2019, Figure 7).

The BSS values in Figure 5.4, where the thresholds correspond to the 40th, 60th, 90th, and 95th percentiles of observed non-zero GHI (25, 127, 498, 604

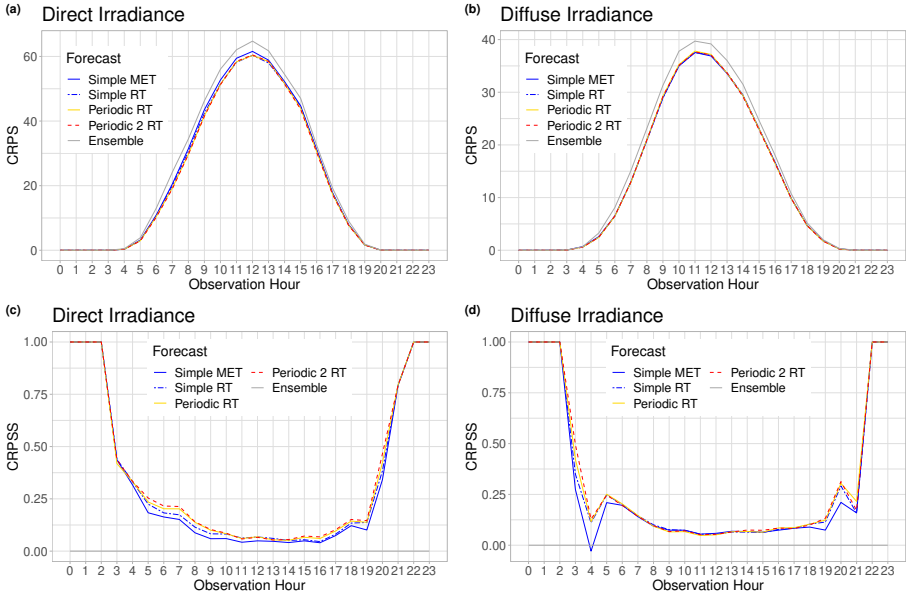

Figure 5.7: Mean CRPS of locally post-processed and raw ensemble forecasts of direct (a) and diffuse (b) irradiance, and corresponding skill scores (c,d) with respect to the raw ensemble as functions of the observation hour for the ICON-EPS dataset.

W/m^2), show a similar pattern. The higher the threshold, the shorter the time with a positive mean BS, as the higher thresholds are often observed around midday, when the irradiance is highest. Post-processed forecasts outperform the raw ensemble for the corresponding lead times, negative skill scores only occur at the boundaries where the mean score values to compare are very small.

The improved calibration of post-processed forecasts is further confirmed by Figure 5.5a, which shows the coverage of the nominal 83.33 % central prediction intervals. When positive global irradiance is likely to be observed between 3 and 19 UTC, the EMOS model produces coverage close to the nominal value, while the raw ensemble's coverage is consistently below 60%.

Furthermore, the MAE of the median forecasts in Figure 5.5b shows that post-processing increases the accuracy of point forecasts as well. The difference in MAE during peak irradiance hours exceeds 20 W/m^2. This is in strong contrast to the findings of the second case study (see Figure 5.14) and shows the existence of a bias in the AROME-EPS that is mitigated by post-processing, as we will see below. The RMSE of the mean forecasts yields similar conclusions (not shown).

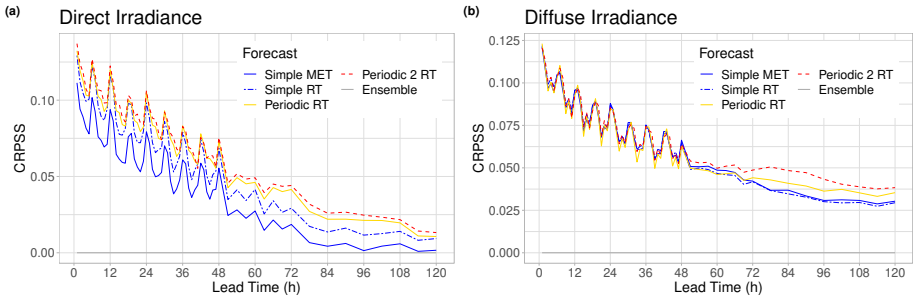

Figure 5.8: CRPSS of locally post-processed forecasts of direct (a) and diffuse (b) irradiance with respect to the raw ensemble as functions of the lead time for the ICON-EPS dataset.

The presence of a bias in the raw ensemble forecasts can also be observed in the verification rank histograms of Figure 5.6. Furthermore, the distinctly U-shaped verification rank histograms reveal a significant underdispersion, which is consistent with the low coverage of the raw ensemble forecasts shown in Figure 5.5a and persists throughout all lead time ranges considered. The ensemble members, on the other hand, are more likely to underestimate the true irradiance, indicating a negative bias. Statistical post-processing effectively corrects

both deficiencies. The upper row of Figure 5.6 shows PIT histograms of EMOS predictive distributions that are almost flat, indicating only a small bias for observations between 12 and 24 UTC.

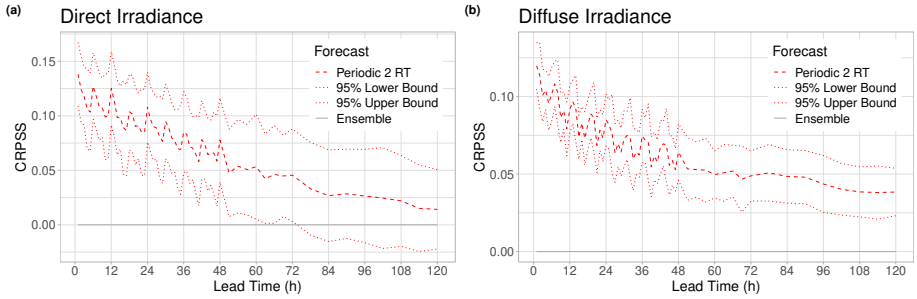

Figure 5.9: CRPSS of the best performing locally post-processed forecasts of direct (a) and (b) diffuse irradiance with respect to the raw ensemble together with 95 % confidence intervals for the ICON-EPS dataset.

5.4.2 Results for the ICON-EPS dataset

The ICON-EPS dataset, in contrast to the first case study, spans a much longer time period, allowing for the consideration and comparison of more complex model formulations and estimation procedures. Remember that we are using forecasts of direct and diffuse irradiance at temporal resolutions of 1 hr (for lead times up to 48 hrs), 3 hrs (for lead times 51 – 72 hrs), and 6 hrs (for lead times 78 – 120 hrs).

Members of the ICON-EPS may be considered exchangeable since they are generated by random perturbations. As a result, for post-processing, we utilize the CL0-EMOS model, with locations linked to ensemble members via (5.7) or (5.8) with $K = 1$. Thus, for model (5.3) with location (5.7) (which was the only model variant considered in Section 5.4.1 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 (5.8) of the location parameter via (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 time period from 27 December 2018 to 31 December 2019 is only used for training, while calendar year 2020 (366 calendar days) is used for model

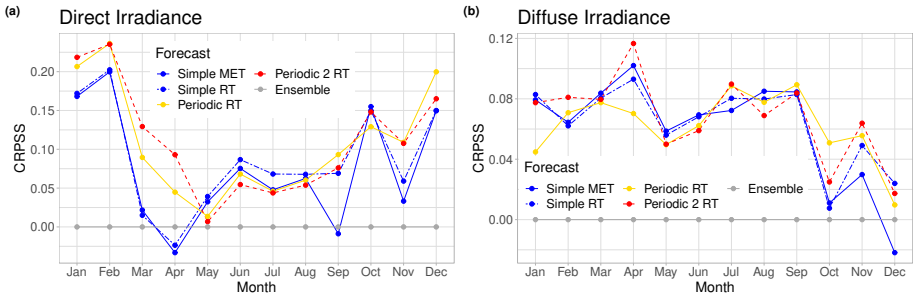

Figure 5.10: CRPSS of locally post-processed forecasts of direct (a) and diffuse (b) irradiance with respect to the raw ensemble, computed based on monthly mean values for the ICON-EPS dataset.

verification, giving enough flexibility to choose a sufficiently long training period even for local modeling. Two training configurations are investigated: a 365-day RTP and a monthly expanding training (MET) scheme, in which all data up to the end of the previous month before the forecast date under consideration is used for training. The first training period in the latter case covers all data prior to calendar year 2020. According to preliminary research (not shown), MET only offers fair verification scores for the simple model. As a result, we present results for the simple model with rolling training (*simple RT*) and monthly expanding training (*simple MET*), as well as the periodic models with rolling training (*periodic RT and periodic 2 RT*).

As reference models, raw ensemble forecasts of direct and diffuse irradiance are used. Unless otherwise stated, the results discussed below are averages of all three observation locations and all four NWP model initialization times. Because of the interacting effects of forecast initialization time, lead time, and corresponding time of day of observation, the interpretation of the results may be more involved than in the first case study.

First, we investigate diurnal effects by analyzing the dependence of the mean CRPS of the various forecast models on the time of the observation shown in Figures 5.7a,b. To provide a reasonable comparison, only lead times up to 48 hrs are considered where hourly forecasts are available. At all time points when positive irradiance is possible, all post-processing methods outperform the raw ensemble forecasts for both direct and diffuse irradiance. According to the skill scores shown in Figure 5.7c, predictive performance in the case of direct irradiance is primarily determined by the complexity of model formulations and

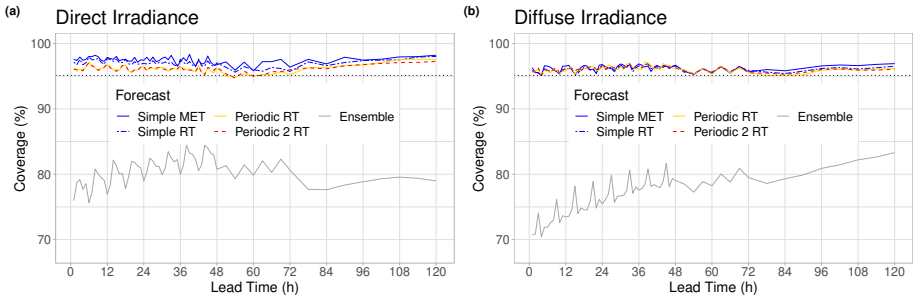

Figure 5.11: Coverage of nominal 95.12 % central prediction intervals of locally post-processed and raw forecasts of direct (a) and diffuse (b) irradiance for the ICON-EPS dataset.

parameter estimation, with more complex models exhibiting better forecast performance. However, the differences between the various EMOS approaches are relatively minor. The same is true for diffuse irradiance in the early and late hours (see Figure 5.7d), while there is no discernible difference in the skill of the different EMOS models between 6 and 18 UTC. It should be noted that the apparent periodic oscillations in the CRPSS values can be caused in part by the pooling of different observation hours as a result of the four considered initialization times. Post-processing, in comparison to the AROME-EPS, improves predictive performance even at night, achieving a CRPS of nearly zero, which is not the case for the raw ICON-EPS.

The observations from Figure 5.7c are confirmed by Figure 5.8a, which shows the CRPSS with respect to the raw direct irradiance ensemble forecasts as a function of lead time. Because of the different scaling of the vertical axis, the variations are more pronounced here, and once again, the periodic 2 model with RTP has the best forecast skill, while the simple model with MET has the smallest CRPSS. In general, longer lead times result in lower skill scores, which is also true for the corresponding CRPSS values for diffuse irradiance (Figure 5.8b). Overall, increases in direct irradiance are slightly greater than in diffuse irradiance, and none of the models result in negative skill scores. There are no discernible distinctions between the different EMOS approaches up to a lead time of 48 hrs. For longer lead times, comparable to direct irradiance, the most complex periodic 2 model performs better, while the simple model with parameters estimated using a rolling training period is now the least skillful. Remember that forecasts with longer lead times apply to a longer time period,

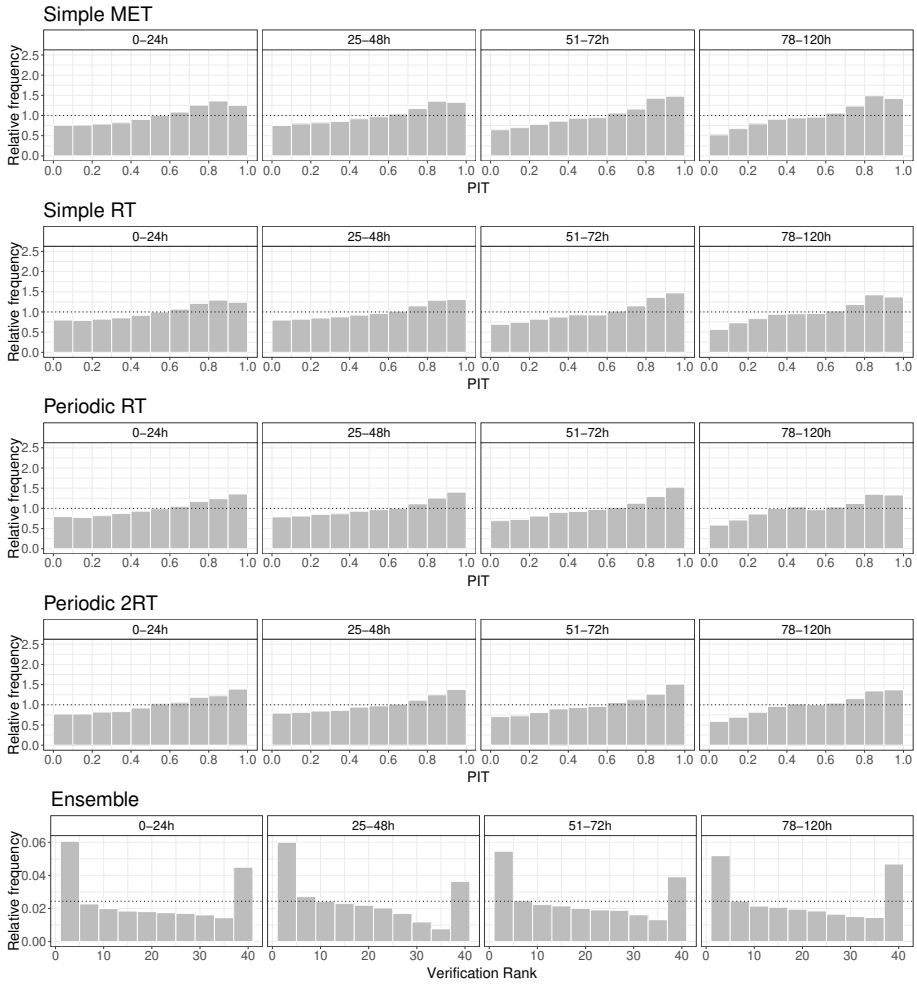

Figure 5.12: PIT histograms of locally post-processed and verification rank histograms of raw ensemble forecasts of DNI for lead times 1-24h, 25-48h, 51-72h and 78-120h for the ICON-EPS dataset.

and thus seasonal effects related to the diurnal cycle can be captured by more complex models.

Figure 5.9 shows the CRPSS of the best performing periodic 2 RT EMOS bootstrap with respect to the raw ICON-EPS forecast and the corresponding 95 % confidence intervals. In the case of direct irradiance the improvement in mean CRPS is significant up to 60 hr, whereas for diffuse irradiance it is significant for all considered lead times. Further, comparing Figures 5.9 and 5.8 it can be observed (especially in the case of diffuse irradiance), that in terms of mean CRPS there is no significant difference between the various post-processing methods.

A third consideration is the forecast skill's reliance on the location of observation. Table 5.1 displays the average CRPSS of the various EMOS models with respect to the raw forecasts, as well as the corresponding CRPSS values of the three different cities for four different lead time intervals. The conclusion from these results is that the magnitude of improvements in predictive performance that result from post-processing is highly dependent on location. For both variables, Karlsruhe gains the most, while the simple MET model performs worse than the raw direct irradiance ensemble forecast and results in negative skill scores for Berlin after 24 hr and Hamburg after 78 hr lead time. Among the competing direct irradiance models, the most complex periodic 2 RT model has the best forecast skill for Berlin and Hamburg, as well as the best overall performance. The variations in performance between the different EMOS models are much smaller in the case of diffuse irradiance, which is consistent with the results shown in Figure 5.8b. None of the more complex models, in particular, consistently outperform the simple MET approach.

Figure 5.10 displays the CRPSS of post-processed forecasts based on monthly mean values to examine seasonal effects in the improvements achieved by post-processing. The enhancement in direct irradiance are generally greater in winter than in summer. From November to April, the variations in post-processing approaches are most pronounced, with more complex model formulations that integrate seasonal effects performing especially well. The overall level of advancement in terms of mean CRPS is lower for diffuse irradiance (mind the different scale of the vertical axes) and there are only mild seasonal effects in the form of smaller corrections between October and December.

To make the results easier to understand, the rest of this section considers combined data from all locations, months, and observation hours, and only shows the dependence on the lead time.

The enhanced calibration of post-processed forecasts can also be seen in Figure 5.11, showing the coverage plots. For all lead times, all post-processing

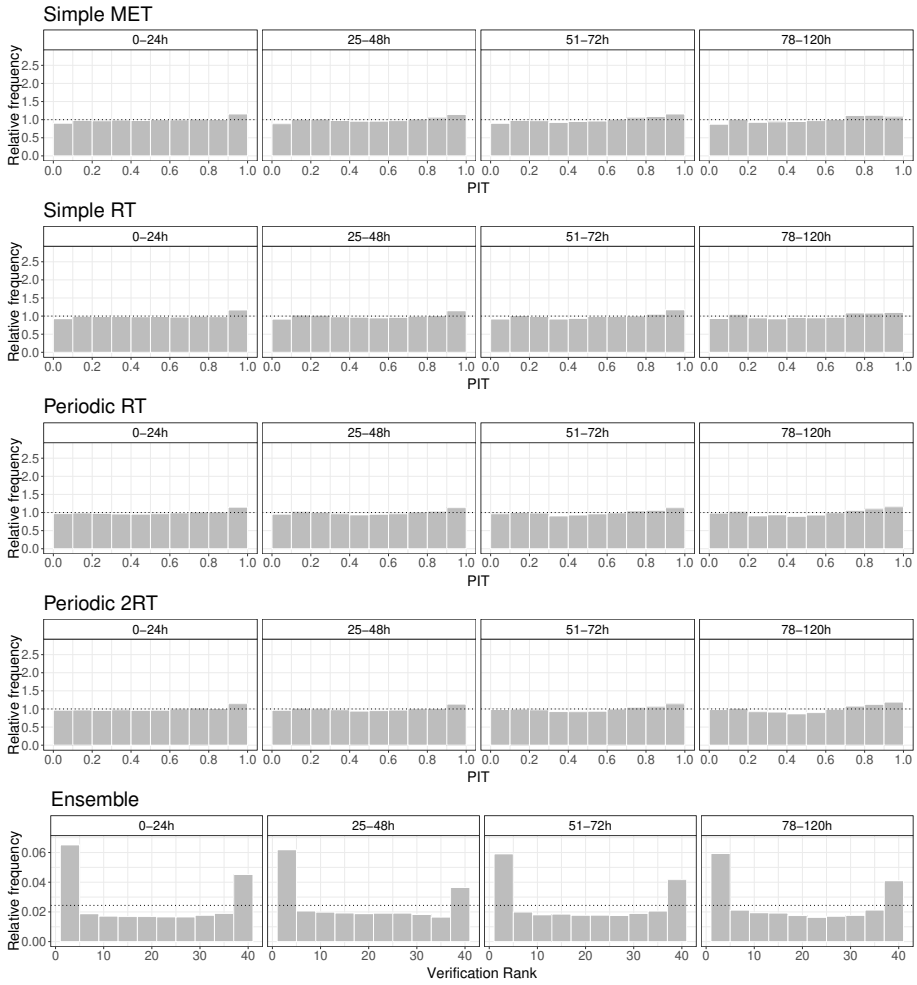

Figure 5.13: PIT histograms of locally post-processed and verification rank histograms of raw ensemble forecasts of DHI for lead times 1-24h, 25-48h, 51-72h and 78-120h for the ICON-EPS dataset.

Direct irradiance

Lead time	Model	Overall	Karlsruhe	Berlin	Hamburg
1-24h	Simple MET	0.076	0.114	0.029	0.082
	Simple RT	0.095	0.130	0.070	0.083
	Periodic RT	0.101	0.133	0.077	0.090
	Periodic 2 RT	0.104	0.132	0.082	0.097
25-48h	Simple MET	0.050	0.091	-0.000	0.054
	Simple RT	0.064	0.102	0.032	0.054
	Periodic RT	0.072	0.102	0.045	0.066
	Periodic 2 RT	0.074	0.099	0.047	0.073
51-72h	Simple MET	0.021	0.062	-0.018	0.018
	Simple RT	0.033	0.071	0.008	0.020
	Periodic RT	0.043	0.075	0.021	0.030
	Periodic 2 RT	0.046	0.069	0.024	0.043
78-120h	Simple MET	0.004	0.028	-0.016	-0.002
	Simple RT	0.013	0.036	-0.001	0.001
	Periodic RT	0.019	0.032	0.017	0.008
	Periodic 2 RT	0.022	0.024	0.021	0.022

Diffuse irradiance

Lead time	Model	Overall	Karlsruhe	Berlin	Hamburg
1-24h	Simple MET	0.089	0.099	0.075	0.093
	Simple RT	0.091	0.100	0.080	0.092
	Periodic RT	0.089	0.099	0.080	0.086
	Periodic 2 RT	0.091	0.097	0.083	0.092
25-48h	Simple MET	0.066	0.076	0.050	0.070
	Simple RT	0.066	0.075	0.053	0.069
	Periodic RT	0.064	0.071	0.055	0.063
	Periodic 2 RT	0.066	0.069	0.060	0.069
51-72h	Simple MET	0.048	0.055	0.041	0.046
	Simple RT	0.046	0.053	0.041	0.043
	Periodic RT	0.047	0.053	0.045	0.041
	Periodic 2 RT	0.051	0.049	0.055	0.049
78-120h	Simple MET	0.032	0.032	0.039	0.025
	Simple RT	0.031	0.031	0.040	0.021
	Periodic RT	0.038	0.037	0.048	0.026
	Periodic 2 RT	0.043	0.036	0.062	0.030

Table 5.1: Overall CRPSS and CRPSS for individual locations of locally post-processed forecasts of direct and diffuse irradiance with respect to the raw ensemble.

methods produce coverage close to the nominal 95.12 %, while the raw ensemble's maximum coverage is below 85 % for both variables. The disparity between post-processed direct irradiance forecasts is more apparent, with periodic models being the most accurate. The shapes of the PIT and verification rank histograms in Figures 5.12 and 5.13 match these results. For all lead times, raw ensemble forecasts of direct irradiance are highly underdispersive and slightly biased. The PIT histograms of all EMOS models are far closer

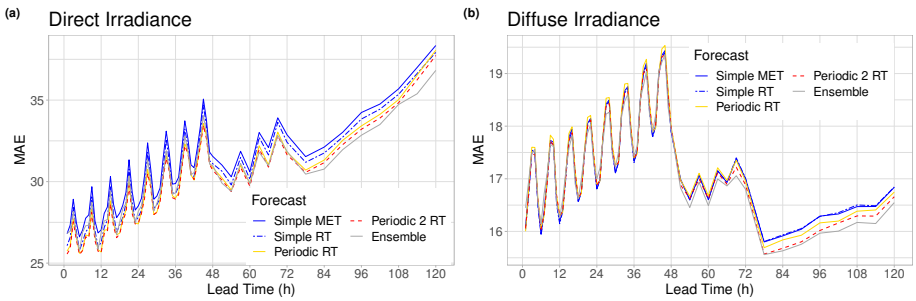

Figure 5.14: MAE of the median forecasts of direct (a) and diffuse (b) irradiance for the various locally post-processed approaches

.

to the optimal uniform distribution, even though this is slightly alleviated for longer lead times. The post-processed forecasts, however, still have some bias.

Neither the PIT histograms of post-processed nor the verification rank histograms of raw diffuse irradiance forecasts show any bias (Figure 5.13), and all EMOS approaches successfully correct the raw ensemble's underdispersion, resulting in nearly perfectly uniform PIT histograms.

Finally, Figure 5.14 showing the MAE of the median forecasts indicates that while post-processing substantially improves the calibration of probabilistic forecasts, it has a minor effect on the accuracy of point forecasts. For all lead times considered, the gap in MAE is less than 2 W/m^2 for direct irradiance and 0.6 W/m^2 for diffuse irradiance. The sharp differences in MAE values at 51 hr and 78 hr are due to a shift in forecast temporal resolution. The RMSE of the mean forecasts results very similar figures, so they are not shown here.

5.5 Conclusions

The proposed post-processing models are capable of improving the forecast performance of raw ensemble predictions reliably and significantly up to lead times of at least 48 hrs. These improvements are larger for AROME-EPS even with the simplest EMOS approach, while for ICON-EPS more complex post-processing models perform better in terms of prediction; however the discrepancies between model variations are rarely statistically significant.

The solar irradiance post-processing models considered in this chapter open

up a number of possibilities for future research. We only used ensemble predictions of the target variable as inputs to the post-processing model in terms of the general model setup. However, the use of modern machine learning methods to integrate arbitrary predictor variables and model potentially nonlinear relations to the forecast distribution parameters has recently become a subject of the post-processing literature (see Vannitsem et al., 2021, for a recent review). The gradient boosting extension method proposed in Messner et al. (2017) or the neural network-based approach proposed in Rasp and Lerch (2018) may be used to expand the EMOS models considered here. See, for example, Sperati et al. (2016) and Bakker et al. (2019) for related considerations in the solar irradiance forecasting literature, where additional predictors from NWP model performance are used, albeit for different types for probabilistic forecasting methods.

In addition, we've focused on univariate forecasts for a single location, lead time, and target variable. Many practical applications, particularly in the context of energy forecasting, require accurate modeling of spatial, temporal, or inter-variable dependencies (Pinson and Messner, 2018). Over the last few years, a wide range of multivariate post-processing methods have been proposed (see Lerch et al., 2020, for a recent overview), and a study of those approaches in the context of solar energy forecasting may be an important starting point for future research.

Finally, the creation of solar irradiance post-processing models was driven by the desire to improve probabilistic solar energy forecasting. To that end, it would be useful to look into the impact of post-processing NWP ensemble forecasts of solar irradiance on PV power prediction, and compare it to direct probabilistic PV power production models (see e.g. Alessandrini et al., 2015). Phipps et al. (2020) finds that a two-phase strategy of post-processing both wind and power ensemble forecasts performs better in a related study in the context of wind energy, and that the calibration of the power predictions is a crucial step. Ideally, statistical post-processing of solar irradiance forecasts will be a key component of modern, fully integrated renewable energy forecasting systems (see e.g. Haupt et al., 2020).

Chapter 6

Total cloud cover prediction using machine learning methods

6.1 Introduction

TCC prediction is critical in observational astronomy (Ye and Chen, 2012) and photovoltaic energy production forecasting, as it is the main cause of variation in solar-radiation energy supply (Matuszko, 2011; McEvoy et al., 2012); however, it is also relevant in agriculture, tourism, and other fields of economy. TCC forecasting in our case can be considered as a nine-group classification problem, requiring methods that are significantly different from those used for other continuous weather variables like temperature, wind speed, or precipitation accumulation.

The main aim of our work here is to investigate the use of machine learning methods for total cloud cover prediction in the context of statistical post-processing of TCC ensemble forecasts, as probabilistic multi-category classification is one of the main areas of application of machine learning. We compare the performance of MLP (Goodfellow et al., 2016), GBM (Friedman, 2000), and RF (Breiman, 2001) with the raw TCC ensemble and the MLR and POLR approaches of Hemri et al. (2016) using ECMWF global ensemble forecasts for the years from 2002 to 2014. We further investigate the effect of using precipitation

ensemble forecasts as additional predictors in TCC post-processing.

In this chapter, first a reference to the calibration methods for TCC forecasts is presented in Section 6.2. Then, we provide a description of the TCC and precipitation ensemble forecasts and observations in Section 6.3, followed by a show case of the results in Section 6.4 and the chapter ends with some conclusions. The following results of this chapter have been published in Baran et al. (2021).

6.2 Post-processing methods

In what follows, let $Y \in \mathcal{Y} = \{y_1, y_2, \ldots, y_9\}$ be TCC at a given location and time expressed in oktas and denote by $\boldsymbol{f} = (f_1, f_2, \ldots, f_{52})$ the corresponding 52-member ECMWF TCC ensemble forecast with a given lead time, where $f_1 = f_{\text{HRES}}$ and $f_2 = f_{\text{CTRL}}$ are the high-resolution and control members, respectively, whereas $f_3, f_4, \ldots, f_{52}$ correspond to the 50 statistically indistinguishable (and thus exchangeable) ensemble members $f_{\text{ENS},1}, f_{\text{ENS},2}, \ldots, f_{\text{ENS},50}$ generated using random perturbations. In this discrete setting the estimation of the predictive distribution of Y reduces to the estimation of conditional probabilities

$$\mathsf{P}\big(Y = y_k \mid \boldsymbol{f}\big), \qquad k = 1, 2, \ldots, 9. \tag{6.1}$$

Obviously, in (6.1) the raw ensemble forecast $\boldsymbol{f}$ can be replaced by any feature vector $\boldsymbol{x}$ derived from the ensemble and/or other covariates. In order to ensure comparability with the reference MLR and POLR approaches (presented in Sections (2.6) and (2.7)) for classification using TCC data only (see Section 6.4.2), we consider the same feature set as in Hemri et al. (2016). For details of the calibration methods we refer to Section 2.4.

The investigated covariates are the HRES forecast f_{HRES}, the control forecast f_{CTRL}, the mean of the 50 exchangeable ensemble members $\overline{f}_{\text{ENS}}$, the ensemble variance

$$s^2 := \frac{1}{51} \sum_{i=1}^{52} \big(f_i - \overline{f}\big)^2, \qquad \text{where} \qquad \overline{f} := \frac{1}{52} \sum_{i=1}^{52} f_i,$$

the proportions of forecasts predicting zero and maximal cloud cover

$$p_0 := \frac{1}{52} \sum_{i=1}^{52} \mathbb{I}_{\{f_i = 0\}} \qquad \text{and} \qquad p_1 := \frac{1}{52} \sum_{i=1}^{52} \mathbb{I}_{\{f_i = 1\}},$$

respectively, and an interaction term

$$I := s^2 \text{sign}(d) d^2 \qquad \text{with} \qquad d := \big((f_{\text{HRES}} - 0.5) + (f_{\text{CTRL}} - 0.5) + (\overline{f}_{\text{ENS}} - 0.5)\big)/3$$

connecting the ensemble variance and the mean deviation of the first three features from 0.5.

As additional feature we also consider the mean $\overline{f}_{\text{PREC}}$ of the ECMWF 51-member precipitation ensemble forecast for some of the models (see Section 6.4.3). The use of the HRES precipitation forecast or of the mean of the 52-member precipitation ensemble (including HRES) instead of $\overline{f}_{\text{PREC}}$ was also tested; however, these models did not result in a significant improvement in the forecast skill.

Verification scores

As mentioned in the Introduction, the case of TCC by forecast F we refer to a discrete probability distribution on $\mathcal{Y}$ characterized by a PMF $p_F(y)$. In this chapter, as proper scoring rules, we going to use CRPS and LogS (defined in sections 3.2 and 3.3) and their corresponding skill scores defined in Section 3.4.

Furthermore, the differences between verification scores is assessed by DM test defined in section 3.9. We address spatial dependencies in simultaneous testing for the different stations using the Benjamini-Hochberg algorithm (Benjamini and Hochberg, 1995) to control the false discovery rate at a 5 % level of significance (see e.g. Wilks (2016)). We further provide confidence intervals for mean score values and skill scores. We also consider the PIT histograms defined in section 3.6. Implementation details for all models are provided in Section 6.4.1.

6.3 Data

We use 52-member ECMWF global ensemble forecasts of TCC and 24 hr precipitation accumulation initialized at 1200 UTC (high-resolution forecast (HRES), control forecast (CTRL), and 50 members (ENS) generated using perturbations in initial conditions (see for instance Buizza (2018b))) for 10 different lead times ranging from 1 day to 10 days for the time interval between 1 January 2002 and 20 March 2014, as well as the corresponding observations. During these 12 years, there were several changes in EPS; however, according to preliminary results (see

e.g. Hemri et al. (2014) and Hemri et al. (2016)), these changes have no significant influence on the skill of the post-processing approaches. The TCC data set is identical to the one investigated in Hemri et al. (2016), which includes data for 3330 SYNOP observation stations which encompasses the entire globe as ECMWF forecasts are issued on the global domain. TCC SYNOP observations are reported in values $\mathcal{Y} = \{0, 0.1, 0.25, 0.4, 0.5, 0.6, 0.75, 0.9, 1\}$ corresponding to the different oktas, whereas the raw ensemble forecasts are continuous values in the $[0, 1]$ interval. The matching of forecasts and observations is performed with quantization of forecast values using intervals

$$[0, 0.01[, \quad [0.01, 0.1875[, \quad [0.1875, 0.3125[, \quad [0.3125, 0.4375[, \quad [0.4375, 0.5625[,$$
$$[0.5625, 0.6875[, \quad [0.6875, 0.8125[, \quad [0.8125, 0.99[, \quad [0.99, 1],$$

that is raw or post-processed forecasts falling e.g. into the interval $[0.1875, 0.3125[$ correspond to observation value 0.25 (see Table A1 in Hemri et al. (2016)).

As quality control, stations having questionable or excessively large amounts of missing data are removed from the data set. The percentage of data points that are equal to the previous 10 data points is the main criterion for removing a station from the data set for both TCC and precipitation (Hemri et al., 2014, 2016). A station is considered unreliable if this number surpasses 20%. Stations having recorded observations outside the range $[0, 1]$ and $[0 \text{ mm}, 1826 \text{ mm}]$ are removed for TCC and precipitation, respectively. The range for precipitation span from the smallest to the largest measurements ever taken on the Earth. After this, our additional precipitation data set, which was investigated in Hemri et al. (2014), includes forecast-observation pairs for 2917 SYNOP stations. TCC and precipitation data are available at 2239 of these stations. Note that the corresponding forecasts are interpolated from the model grid using bilinear and nearest-neighbour interpolation schemes.

6.4 Results

All of the calibration methods described in sections 6.2 and 2.4 require training data that is large enough to provide numerical stability and fair predictive performance. Following Hemri et al. (2014), we concentrate on local calibration, that is, post-processing of forecasts for a single station using only that station's training data. As a consequence, in order to obtain a sufficiently large training set, relatively long training periods are required. As in Hemri et al. (2016), we

use 5-year training periods and both non-seasonal and seasonal training schemes to ensure comparability with the reference approaches. Forecasts and observations from the previous five calendar years (e.g., 1 January 2003 to 31 December 2007) are used to train the model for calibration of TCC ensemble forecasts for the entire following calendar year (1 January to 31 December 2008), then the training period is rolled forward by one year (1 January 2004 – 31 December 2008). In the seasonal approach, two different seasons are considered covering April–September and October–March respectively, and the TCC ensemble forecast for a given day is calibrated using only training data from the same season. Because of the use of 5-year training periods, predictive PMFs are available for the period 1 January 2007 to 20 March 2014 (2636 calendar days), where one can measure the forecast skill of the post-processing methods presented in section 6.2.

Furthermore, as Hemri et al. (2016) suggest, numerical problems with LogS calculation can be prevented by replacing unrealistically low values $p_F(y_j)$ of the predictive PMF corresponding to okta y_j with a probability $p_{\min}$ ensuring that okta y_j is observed at least once during the training period with a 1% chance. Translated to formulae, this means that instead of $p_F(y_j)$ we consider $\max\{p_{\min}, p_F(y_j)\}$, where $p_{\min}$ solves $0.01 = 1 - (1 - p_{\min})^T$ with T being the length of the training period in days, and adjust the probabilities to get a PMF again (for more details see Hemri et al. (2016)). Note that this is a minor technical adjustment that results in negligible changes in CRPS or PIT values as compared to the original predictive PMFs.

6.4.1 Implementation details

Here, we discuss implementation details for the different statistical and machine learning methods for TCC post-processing.

Multiclass and proportional odds logistic regression

Both models have a variety of different implementations. The coefficients of different MLR and POLR models are estimated here using respectively the `R` packages `nnet` and `MASS` (Venables and Ripley, 2002). It should be noted that the implementation based on the `nnet` package uses neural networks to estimate the parametric MLR model (2.6), which is a fundamentally different use of neural networks compared to our MLP models presented in Section 2.4.1.

Multilayer perceptron neural networks

In our computations, we use `Matlab`'s `patternnet` function with two hidden layers of 10 and 15 neurons. Both hidden layers use the hyperbolic tangent as activation function. We use the LogS loss function (also known as cross-entropy in machine learning) with a regularization parameter of 0.1 and scaled conjugate gradient as minimization algorithm. The corresponding data set is split into a training and validation set for each 5-year training period (both for seasonal and non-seasonal approaches), the latter being a randomly selected subset consisting of 15 % of the data. Training with a growing data set using all available forecast cases from previous years while simultaneously increasing the weight of the regularization term was also tested as an alternative to the 5-year RTP. However, this approach did not result in an improved forecast skill.

Random forests

Our RF model implementation is based on the `R` package `XGBoost` (Chen et al., 2019). For a particular observation station and forecast horizon, the tuning parameters (depth of trees, number of predictors sub-sampled at each splitting node) are determined as follows. The first of the rolling 5-year training periods, from 2002 to 2006, is divided into an initial training set (2002–2005) and a validation set (year 2006). RF models consisting of 300 trees are estimated based on the initial training set and evaluated on the validation set using the LogS for all combinations of tree depths between 2 and 4 and number of predictors between 1 and 3. The tuning parameters that result in the lowest LogS on the validation set are then used to fit a 1000-tree RF model for the entire training set (years 2002 to 2005) and to generate forecasts for the first out-of-sample test set (year 2007). This optimal combination of tuning parameters is also used for all subsequent 5-year training periods for that particular station and lead time to reduce computational costs.

Tree depths of 2, 3, and 4 are selected in about 43 %, 36 % and 21 % of cases, respectively, for 5-year RTPs. The number of predictors chosen for subsampling is slightly more uniformly distributed, with trees of depth 3 with 3 predictors sub-sampled at each split (around 17 % of all cases) being the most commonly chosen tuning parameter combination. We did not consider a larger set of possible parameter values to reduce computational costs because initial testing did not show improvements in predictive performance and RF models are also reasonably robust to the choice of tuning parameters.

Gradient boosting machines

We use the `R` package `XGBoost` (Chen et al., 2019) to implement GBM models. Throughout, we use shrinkage with a learning rate of $\lambda = 0.1$, to reduce the effect of each individual tree hm^c by only adding a scaled version of that tree. We use an early stopping criterion to determine the number of boosting iterations M for a given tree depth to further prevent overfitting. To that end, each 5-year training set is divided into two parts: an initial training set (the first four years) and a validation set (the last year). GBM models of the form (2.8) are then estimated iteratively for $m = 1, 2, \ldots$, based on the initial training set until the LogS on the validation set has not improved during the last 25 iterations. This process is repeated for all tree depth values between 1 and 4, and the combination of tree depth and the corresponding optimal number of boosting iterations that produces the best LogS on the training set is chosen as a set of tuning parameters. The final out-of-sample forecasts for the test set are produced based on a GBM model fitted on the full training set using these tuning parameters. A separate set of tuning parameters is determined according to the procedure described above for any combination of station and lead time, and any of the 5-year RTP.

For models with a 5-year RTP, an optimal tree depth of 1 is chosen in approximately 86.5 % of all GBM models, a depth of 2 in approximately 11.5 % of the cases, and a depth of 3 or 4 in less than 2 % of the cases. The average number of boosting iterations is 78.3, but this varies widely depending on the depth of the corresponding tree.

When fitting seasonal RF and GBM models, the procedures for determining optimal tuning parameters of RF and GBM models mentioned above are applied separately to the two seasons. As a result, the sets of optimal tuning parameters for those variants differ not only by station, lead time, and year (only for GBM), but also by season.

6.4.2 Post-processing of TCC ensemble forecasts

As a first step, we look at the use of the MLP, RF, and GBM approaches to post-process TCC ensemble forecasts. The raw TCC ensemble forecast, as well as the MLR and POLR models, are considered as references. All calibrated forecasts are based on the 7-dimensional feature vector $\left(\overline{f}_{\mathrm{ENS}}, f_{\mathrm{CTRL}}, f_{\mathrm{HRES}}, s^2, p_0, p_1, I\right)^\top$ except the MLR, where following Hemri et al. (2016) the number of parameters is reduced by omitting the interaction term I. It should be noted that the MLP model was also tested with the

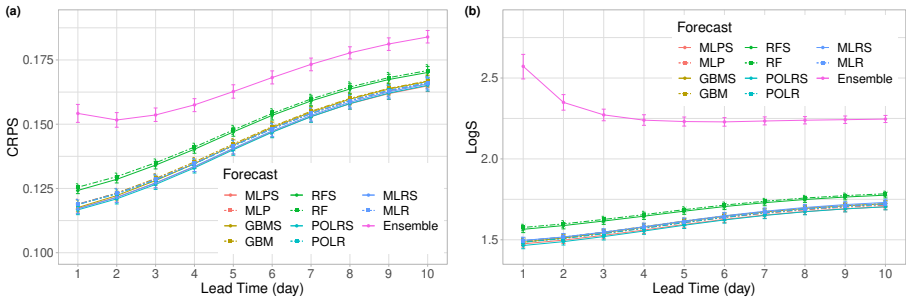

Figure 6.1: Mean CRPS (a) and LogS (b) of the raw ensemble and post-processed forecasts together with 95 % confidence intervals.

52-member TCC forecast ensemble as a feature vector, but this approach did not improve predictive performance. Further, in the POLR model, the coefficients of $\overline{f}_{\mathrm{ENS}}$, f_{CTRL} and f_{HRES} are forced to be non-negative by iterative exclusion of covariates with negative weights, following the suggestions of Hemri et al. (2016). Finally, we test both non-seasonal and seasonal training for all five calibration approaches. Forecasts obtained using the latter are referred to as MLPS, RFS, GBMS, MLRS, and POLRS, respectively.

The mean CRPS and LogS of the raw ensemble and post-processed forecasts, as well as 95 % confidence intervals as functions of the lead time, are shown in Figure 6.1. All calibrated TCC forecasts outperform the raw ensemble by a large margin, and the different approaches are clearly grouped. The MLP, GBM, POLR, and MLR methods, as well as their seasonally estimated counterparts, produce the lowest mean CRPS and LogS values and display very small differences in forecast skill. The non-seasonally and seasonally estimated RF forecasts are in the second group, with the latter yielding slightly lower score values than the former.

Figure 6.2 plotting the CRPSS and LogSS with respect to the POLRS forecasts, which shows the best forecast skill among the methods studied in Hemri et al. (2016), makes it easier to compare the performance of the forecasts in the first group. According to Figure 6.2a, POLRS outperforms its competitors in terms of mean CRPS up to day 7, while MLPS has the best predictive performance for longer lead times. Forecasts based on seasonal training produce lower mean CRPS than non-seasonal counterparts in general, but the differences decrease as the lead time increases. Results in terms of the LogS shown in Figure

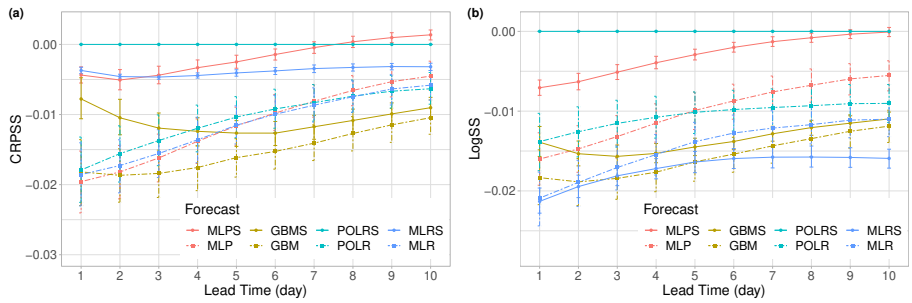

Figure 6.2: CRPSS (a) and LogSS (b) with respect to the POLRS model of MLPS, MLP, GBMS, GBM, MLRS, MLR and POLR forecasts together with 95 % confidence intervals.

6.2b indicate a different behavior and ranking of the models namely the mean LogS of the MLPS approach reaches that of the POLRS model only at day 10 and MLRS underperforms all other methods for all lead times.

These observations are further supported by Figure 6.3 showing the proportion of stations where DM test indicates significant difference in mean CRPS and LogS for lead times 1, 4, 7 and 10 days. To simplify the presentation we compare only the raw ensemble and seasonally trained versions of the calibration approaches, as seasonal models in general outperform their non-seasonal counterparts. The raw ensemble and RFS forecasts are distinctly separated from the other four methods for all lead times, as most of the corresponding cells' entries are almost 100 %. For longer lead times, GBMS outperforms its rivals in almost all stations, both in terms of CRPS and LogS. On the other hand, as the lead time is increased, the proportion of stations where the mean LogS of MLPS and POLRS forecast decreases, whereas in terms of the mean CRPS after decrease one can observe a slight increase. This behaviour is in line with the MLPS skill scores of Figures 6.2a and 6.2b, respectively. Overall, we find that, despite the fact that the absolute differences in CRPS and LogS between the different approaches are small, they thus are often statistically significant for a large proportion of the stations.

Figure 6.4 demonstrates the positive effect of post-processing in the PIT histograms, where only the results for better performing seasonally trained models are reported. The raw ensemble's U-shaped histograms at days 1 and 4 clearly show underdispersion, while a slight hump appears at days 7 and 10. For short lead times, RFS forecasts are overdispersive, and as the forecast horizon in-

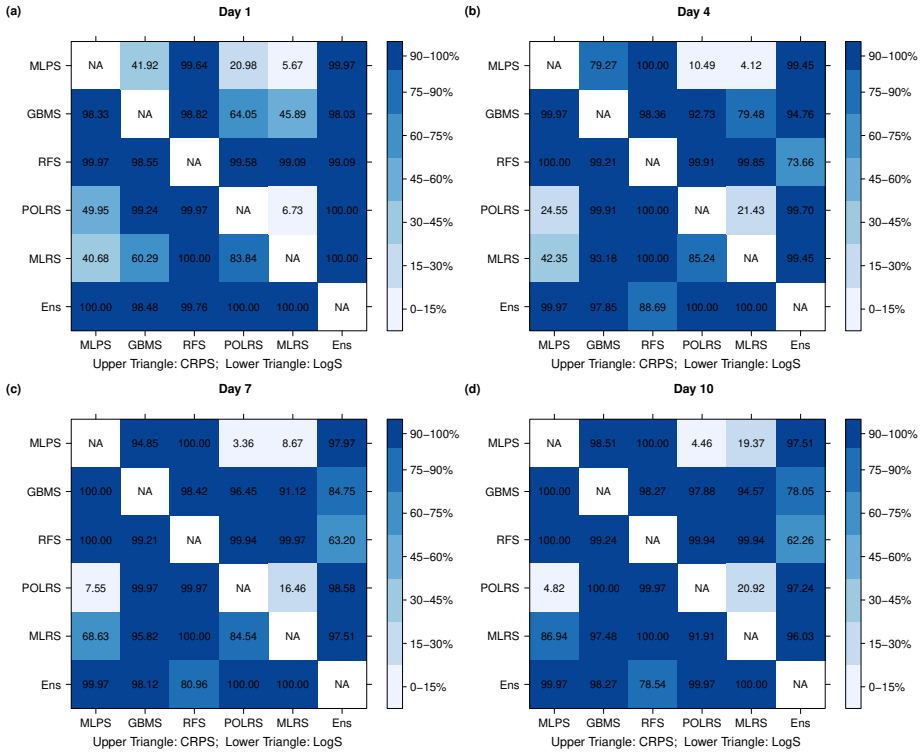

Figure 6.3: Proportion of stations with significantly different mean CRPS (upper triangle) and LogS (lower triangle) at a 5 % level of significance for lead times 1 (a), 4 (b), 7 (c) and 10 (d) days.

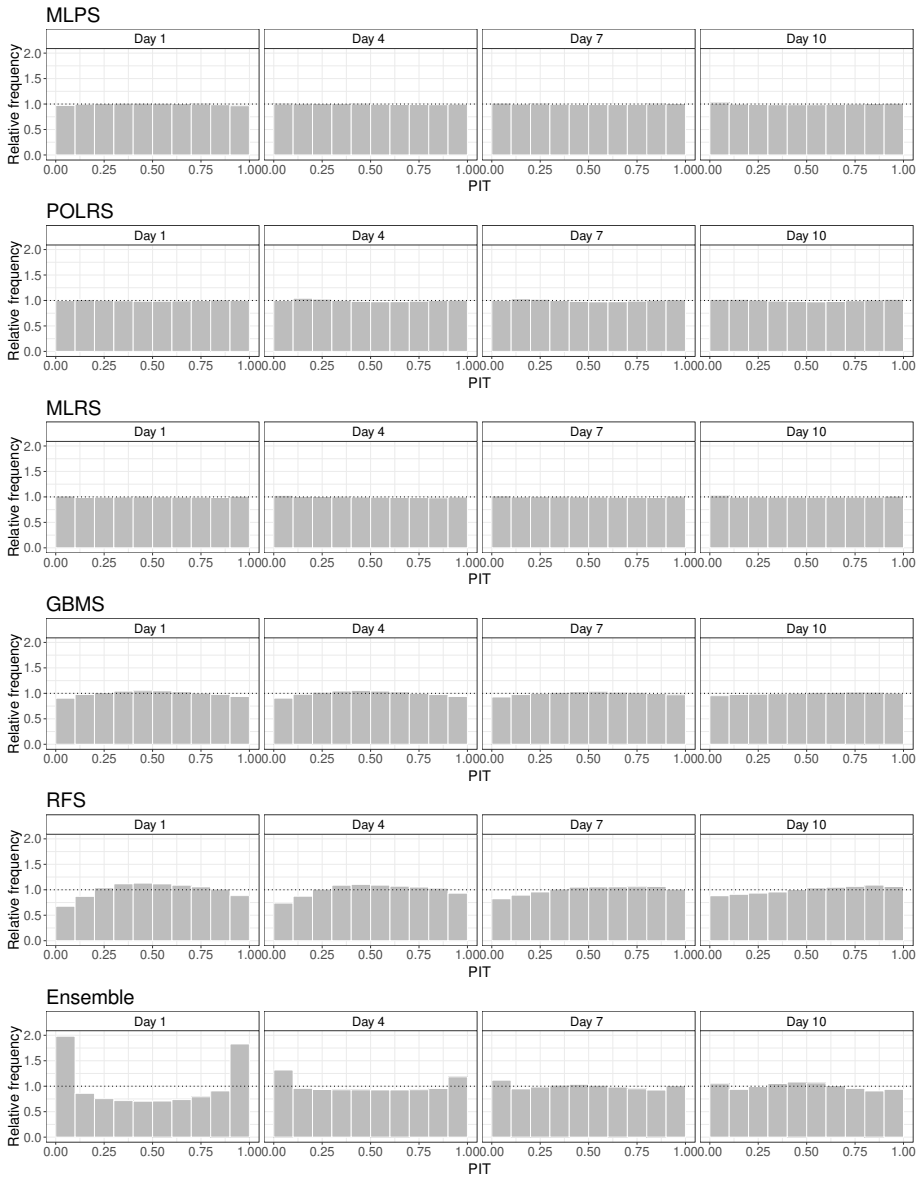

Figure 6.4: PIT histograms over all stations and dates (3300 stations, 2636 days) of the seasonally trained calibration approaches and the raw ensemble at days 1, 4, 7 and 10.

creases, some bias emerges. GBMS forecasts exhibit the same behaviour, however, to a much smaller extent. POLRS and MLPS PIT histograms are almost perfectly flat, suggesting better calibration than the other approaches.

6.4.3 Post-processing using an extended feature set

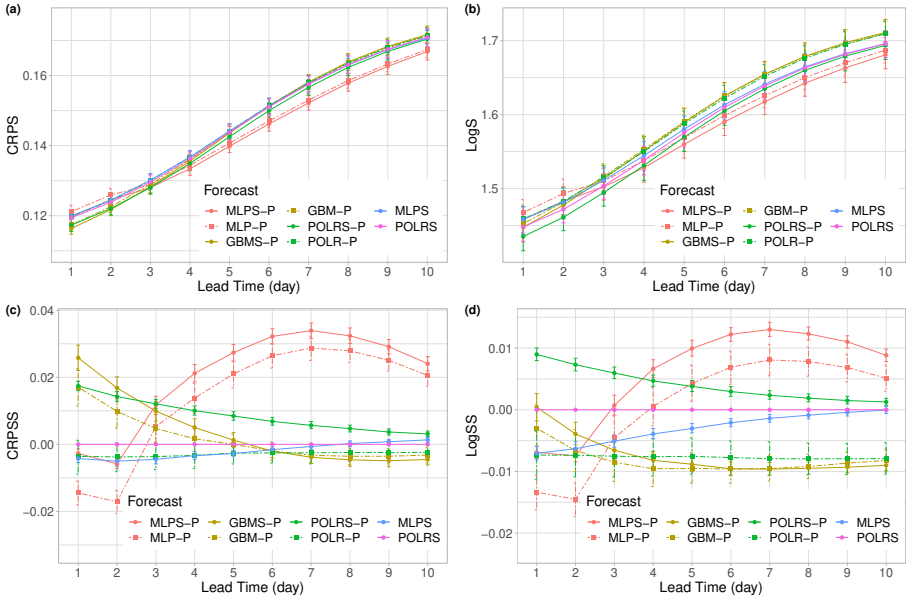

Figure 6.5: CRPS (a) and LogS (b) of different MLP, GBM and POLR forecasts and the corresponding skill scores with respect to the POLRS model (c,d) together with 95 % confidence intervals.

Several recent studies on post-processing (e.g. Taillardat et al. (2016); Rasp and Lerch (2018); Bakker et al. (2019)) have shown the added value of incorporating additional features based on geographical data from SYNOP stations and/or forecasts of other weather variables. Functionals of precipitation ensemble forecasts are a natural option for additional predictors because of their direct connection to clouds (Mishra, 2018). We here use the mean $\overline{f}_{\text{PREC}}$ of the ECMWF 51-member precipitation forecast as additional covariate and investigate the performance of MLP, GBM and POLR approaches, showing the

best forecast skill in Section 6.4.2, with extended feature vector

$$\left(\overline{f}_{\text{ENS}}, f_{\text{CTRL}}, f_{\text{HRES}}, s^2, p_0, p_1, I, \overline{f}_{\text{PREC}}\right)^\top.$$

We take into account both non-seasonal and seasonal training, and the corresponding models are reported as MLP-P, GBM-P, POLR-P, and MLPS-P, GBMS-P, POLRS-P, respectively.

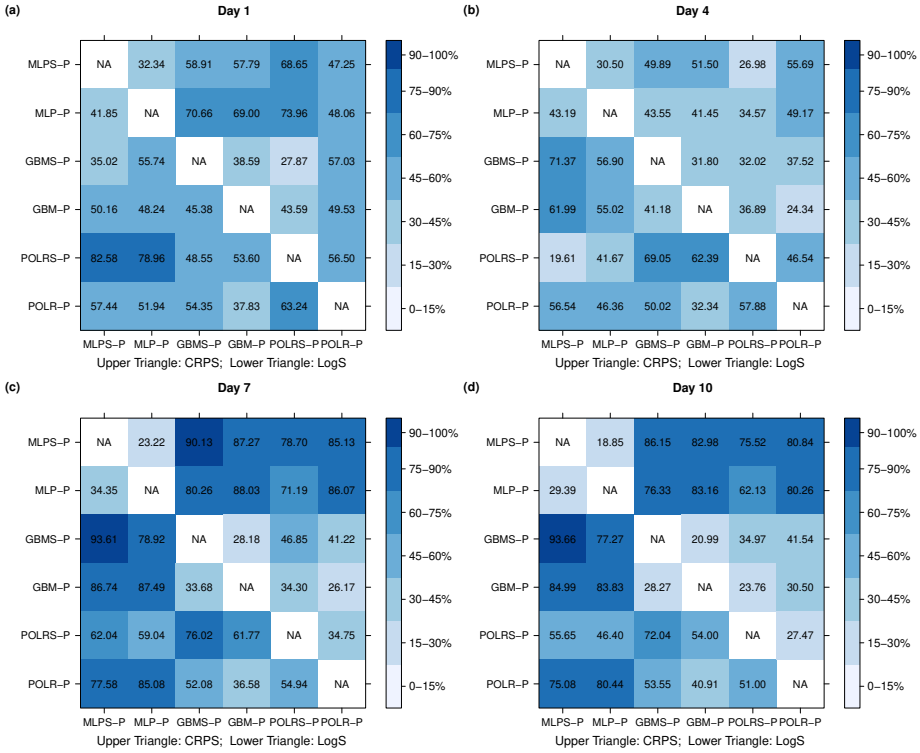

Figure 6.6: Proportion of stations with significantly different mean CRPS (upper triangle) and LogS (lower triangle) at a 5 % level of significance for lead times 1 (a), 4 (b), 7 (c) and 10 (d) days.

According to Figures 6.5a and 6.5b, where the mean CRPS and LogS values of different MLP, GBM and POLR forecasts are plotted as functions of the lead time, and Figures 6.5c and 6.5d displaying the corresponding skill scores

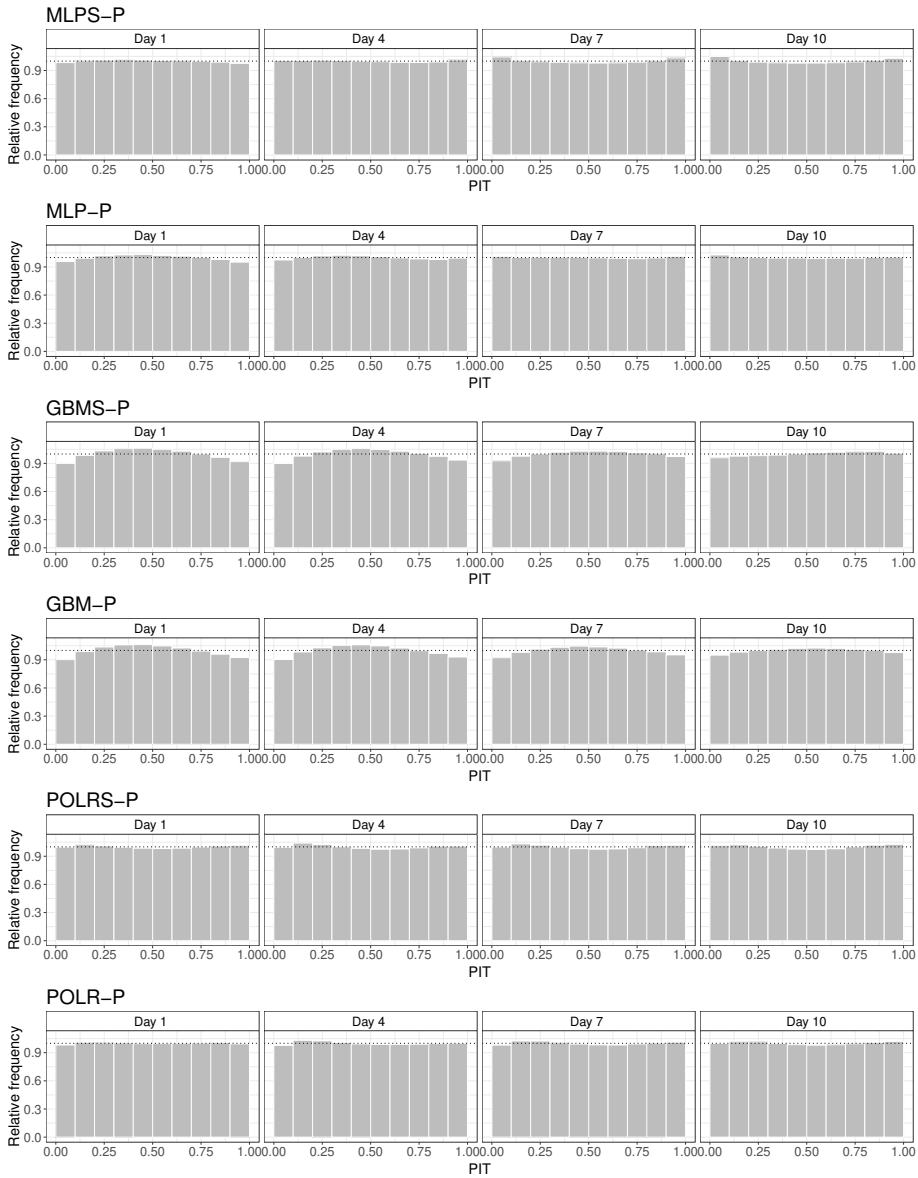

Figure 6.7: PIT histograms over all stations and dates (2239 stations, 2636 days) of the calibration approaches using precipitation forecasts at days 1, 4, 7 and 10.

with respect to the POLRS model, the additional covariate results in different effects for the MLP, and the GBM and POLR models. MLP models that use precipitation forecasts outperform MLP models that only use TCC forecasts after day 2 in terms of both CRPS and LogS, regardless of the training scheme (MLP is not shown), and MLPS-P and MLP-P result in the best predictive performance for longer lead times. The use of precipitation, on the other hand, has the highest effect on POLR models at day 1, and the differences between POLRS-P and POLRS and POLR-P and POLR models (POLR is not shown) decrease as the lead time increases. The GBMS and GBM models exhibit the same phenomenon (not shown). The use of a precipitation forecast significantly improves predictive performance, but the difference decreases as the lead time increases. GBMS-P and GBM-P approaches have lower mean CRPS than the POLRS model up to day 5, but GBMS-P outperforms POLRS-P and MLPS-P for days 1 and 2.

These results are consistent with the proportions of stations with significantly different mean CRPS and LogS values shown in Figure 6.6, where for visual clarity we only consider models with the extended feature set. For example, the proportion of stations where the mean CRPS of the GBMS-P and GBM-P models differ shows a monotone decreasing sequence of 38.59 %, 31.80 %, 28.18 %, 20.99 %, which mimics the decreasing distance of the corresponding curves in Figure 6.5c, while the bow of the CRPSS of MLPS-P with respect to POLRS and the decrease of the CRPSS of POLRS-P matches the change of the corresponding entries (68.65 %, 26.98 %, 78.70 %, 75.72 %) in Figure 6.6.

In terms of calibration, Figure 6.7 shows the PIT histograms of the calibration approaches using precipitation forecasts for days 1, 4, 7 and 10. Generally speaking, all six methods produce reasonably well calibrated predictive PMFs in all considered forecast horizons. For all lead times, the histograms of the GBMS-P and GBM-P approaches are overdispersive, while the histograms of the MLPS-P, MLP-P, and POLR-P approaches are slightly overconfident only on day 1, transforming to a small underdispersion at longer lead times. Notice that, unlike Figure 6.4, which uses verification data from 3330 locations, we only use PIT values from 2239 SYNOP stations where precipitation ensemble forecasts are available. However, since the PIT histograms of the raw ensemble and the MLPS, GBMS, and POLRS forecasts do not change in shape as a result of this reduction, they are not shown in this case. Finally, the MLPS-P, MLP-P, GBMS-P, GBM-P, POLRS-P, and POLR-P approaches almost completely inherit the general behavior of the MLPS, MLP, GBMS, GBM, POLRS, and POLR forecasts in terms of PIT values.

6.5 Conclusions

In this chapter, we propose a novel method for statistical post-processing of TCC ensemble forecasts using a variety of machine learning based classification techniques. This is the first study to compare cutting-edge machine learning approaches with parametric classification methods. The superiority of neural network classification over the best parametric models is demonstrated in an extended case study focused on the ECMWF global TCC ensemble forecasts for 2002–2014 (Hemri et al., 2016). The possibility of incorporating further covariates into TCC statistical post-processing is also investigated. When the mean precipitation accumulation is applied as additional covariate, the results indicate that MLP neural network classification has the best predictive performance for long lead times. The flexibility of neural network models, as well as the broad variety of reasonable covariates, opens the door to further research. These studies may have a direct economic benefit since more accurate TCC prediction is important in many industries, such as PV energy production, agriculture, and tourism. Several of the probabilistic classification methods exhibit complementary systematic errors in calibration. As a result, forecast combination techniques like the ones investigated in Baran and Lerch (2018) might be able to boost predictive performance. The related topic of calibrating and combining probabilistic classifiers has recently sparked interest in the machine learning literature, see e.g. Kull et al. (2019).

The MLP model, in particular, offers some interesting starting points for potential extensions due to the flexibility of neural network model architectures. Long short-term memory neural networks (Hochreiter and Schmidhuber, 1997), for example, are commonly used for time series modeling and may enable temporal dependencies of forecast errors of raw ensemble predictions to be incorporated. Furthermore, techniques similar to those suggested in Rasp and Lerch (2018) could theoretically aid in the construction of a single MLP model jointly for all stations that is still locally adaptive.

As precipitation data is added, the performance improves even further compared to the one without adding an extended feature set, demonstrating the benefits of modern machine learning approaches like GBM and MLP for total cloud cover prediction. In comparison to the traditional MLR and POLR models, these methods make it simple to incorporate new predictors and provide tools for avoiding overfitting. Additional predictor variables, such as indices of atmospheric stability, pressure, humidity, and temperature information at higher levels of the atmosphere, or seasonal information, can boost predictive performance even more. Furthermore, using techniques such as measures of fea-

ture importance (Rasp and Lerch, 2018; Breiman, 2001), more complex machine learning models incorporating several predictors which not only boost TCC predictions, but also allow for a better understanding of the shortcomings of the raw ensemble predictions.

Summary

The thesis presents several statistical and machine learning post-processing techniques for various weather and hydrological quantities. We consider water level, solar irradiance and total cloud cover, and propose novel post-processing methods for each of these variables. The efficiency of the suggested approaches is tested in several case studies, the ensemble forecasts and observations used in this research were produced by different ensemble prediction systems, cover various geographical areas and the temporal resolutions also vary.

Water level forecasting

We present a new BMA model for calibrating Box-Cox transformed hydrological ensemble forecasts of water level, which produces a predictive distribution that is a weighted mixture of doubly truncated normal distributions. The model with three different parameter estimation approaches is evaluated on the 79 member ensemble forecast of BfG for water level at gauge Kaub of the Rhine, for 120 different lead times. The CRPS of the probabilistic forecast distributions and the MAE of the corresponding median forecasts are used for verification. We also took a look at coverage and the average width of nominal central prediction intervals, which is a sharpness metric. The forecast skill of the BMA model is also compared to that of the raw ensemble and the recently implemented EMOS model of Hemri and Klein (2017).

Based on our findings, one may infer that, compared to the raw ensemble, post-processing often improves probabilistic calibration and point forecast accuracy. Furthermore, the BMA model, which uses pure ML for parameter estimation, has the best predictive performance, and, with the exception of very short lead times, it outperforms the EMOS calibration significantly. A direct comparison of the CRPSS values obtained in the case of water level forecasting

(shown in Figure 4.5a) shows that seasonal and analog dependent training periods outperform RTP at least for EMOS. Hence, this suggests that using a more advanced post-processing method or a more intelligent training period selection is fairly unnecessary. As a result, if a sufficiently long set of hydrological hindcasts is accessible, we suggest using EMOS with analog-based training periods and BMA otherwise.

Solar energy forecasting

We suggest a post-processing approach for solar irradiance ensemble weather predictions in which probabilistic forecasts are obtained as a logistic distribution left-censored at zero. In two case studies, several model variants are evaluated that vary in terms of the temporal composition of training datasets and changes to seasonal variations in the model formulation. Despite the fact that the case studies cover a variety of geographical areas, NWP systems, solar irradiance types, and temporal resolutions, the results in Section 5.4 show that the proposed post-processing models can reliably and significantly boost the forecast performance of the raw ensemble predictions up to lead times of at least 48 hrs. The AROME-EPS dataset benefits more from post-processing, which may be due to a lower skill of the raw ensemble predictions due to a bias in addition to the observed underdispersion. We found that more complex post-processing models have better predictive performance on the ICON-EPS dataset, but the discrepancies between model variations are rarely statistically significant.

In the case of the ICON-EPS dataset, the overall level of improvements achieved through statistical post-processing of the raw ensemble's solar irradiance forecasts are comparable to meteorological variables such as precipitation accumulation (Scheuerer, 2013; Baran and Nemoda, 2016) or total cloud cover (Baran et al., 2021), and these enhancements are even slightly larger for the AROME-EPS results. Post-processing ensemble predictions of such variables is often seen as a more difficult task when compared to variables like temperature (Gneiting et al., 2005) or wind speed (Thorarinsdottir and Gneiting, 2010), where significant improvements can be made. Nonetheless, for lead times of up to 2 days, the observed improvements are statistically significant, and will likely be relevant for solar energy forecasting in terms of possible economic benefits and enhanced demand and supply balancing for incorporating dynamic PV power systems into the electrical grid.

Total cloud cover forecasting

For statistical post-processing of total cloud cover ensemble forecasts, we investigate a variety of machine learning classifiers. In particular, we consider multilayer perceptron neural networks, random forest methods and gradient boosting machines, which are tested on ECMWF global TCC ensemble forecasts with lead times from 1 to 10 days and the corresponding discrete SYNOP observations. As reference models, we use raw TCC ensemble forecasts, multiclass and proportional odds logistic regression, and we consider both seasonal and non-seasonal training (following Hemri et al. (2016)).

First we investigate the settings of Hemri et al. (2016), where the classification is solely dependent on predictors determined from the TCC ensemble forecasts. In general, for all lead times, all post-processing methods outperform the raw ensemble in terms of mean CRPS and mean LogS over the verification data, and the corresponding PIT histograms are closer to the uniform distribution than the rank histograms of the raw forecasts. Seasonally trained models further perform slightly better than their non-seasonal counterparts in terms of predictive performance. While the differences between MLP, GBM, POLR, and MLR approaches are generally small, RF models underperform their competitors. The POLR model with seasonal training appears to be the most skillful for short and medium forecast horizons, followed by the seasonally trained MLP model, which performs best for long lead times.

The inclusion of mean precipitation accumulation as a covariate increases predictive performance and alters the ranking of the various methods. The seasonal POLR model outperforms the seasonally and non-seasonally trained MLP with this extended feature set only for short lead times; after days 3–4, it is substantially outperformed by both the seasonally and non-seasonally trained MLP. However, as the lead time increase, the benefit of the extended set of covariates diminishes.

List of publications

Conferences

Baran Sándor, Hemri Stephan, **El Ayari Mehrez**. Statistical post-processing of hydrological forecasts using Bayesian model averaging. *IX. International Workshop on Applied Probability (IWAP 2018)*, Budapest, Hungary, June 18 – 21, 2018.

Baran Sándor, Hemri Stephan, **El Ayari Mehrez**. Statistical post-processing of hydrological forecasts using Bayesian Model Averaging. *The General Assembly of the European Geosciences Union 2019 (EGU2019)*, Vienna, Austria, April 7 – 12, 2019.

Accepted abstract for the German Probability and Statistics Days 2020 (GPSD2020).

Baran Ágnes, Lerch Sebastian, **El Ayari Mehrez**, Baran Sándor. Statistical post-processing of total cloud cover ensemble Forecasts. *Conference on Information Technology and Data Science, CITDS 2020*, Debrecen, Hungary, December November 6–8, 2020. (Online Conference)

UNIVERSITY of DEBRECEN

UNIVERSITY AND NATIONAL LIBRARY
UNIVERSITY OF DEBRECEN
H-4002 Egyetem tér 1, Debrecen
Phone: +3652/410-443, email: publikaciok@lib.unideb.hu

Registry number: DEENK/211/2022.PL
Subject: PhD Publication List

Candidate: Mehrez El Ayari
Doctoral School: Doctoral School of Mathematical and Computational Sciences
MTMT ID: 10077448

List of publications related to the dissertation

Foreign language scientific articles in international journals (3)

1. Baran, Á., Lerch, S., **El Ayari, M.**, Baran, S.: Machine learning for total cloud cover prediction.
   Neural Comput. Appl. 33, 2605-2620, 2021. ISSN: 0941-0643.
   DOI: http://dx.doi.org/10.1007/s00521-020-05139-4
   IF: 5.606 (2020)

2. Schulz, B., **El Ayari, M.**, Lerch, S., Baran, S.: Post-processing numerical weather prediction ensembles for probabilistic solar irradiance forecasting.
   Sol. Energy. 220, 1016-1031, 2021. ISSN: 0038-092X.
   DOI: http://dx.doi.org/10.1016/j.solener.2021.03.023
   IF: 5.742 (2020)

3. Baran, S., Hemri, S., **El Ayari, M.**: Statistical Postprocessing of Water Level Forecasts Using Bayesian Model Averaging With Doubly Truncated Normal Components.
   Water Resour. Res. 55 (5), 3997-4013, 2019. ISSN: 0043-1397.
   DOI: http://dx.doi.org/10.1029/2018WR024028
   IF: 4.309

Total IF of journals (all publications): 15,657
Total IF of journals (publications related to the dissertation): 15,657

The Candidate's publication data submitted to the iDEa Tudóstér have been validated by DEENK on the basis of the Journal Citation Report (Impact Factor) database.

14 April, 2022

DEBRECENI EGYETEM * Egyetemi és Nemzeti Könyvtár *

Bibliography

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.

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.

Baran, Á., Lerch, S., El Ayari, M. and Baran, S. (2021). Machine learning for total cloud cover prediction. *Neural Computing and Applications*, 33, 2605–2620.

Baran, S. (2014). Probabilistic wind speed forecasting using Bayesian model averaging with truncated normal components. *Computational Statistics & Data Analysis*, 75, 227–238.

Baran, S., Hemri, S. and El Ayari, M. (2019). Statistical postprocessing of water level forecasts using bayesian model averaging with doubly truncated normal components. *Water Resources Research*, 55, 3997–4013.

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, 2289–2299.

Baran, S. and Lerch, S. (2016). Mixture EMOS model for calibrating ensemble forecasts of wind speed. *Environmetrics*, 27, 116–130.

Baran, S. and Lerch, S. (2018). Combining predictive distributions for the statistical post-processing of ensemble forecasts. *International Journal of Forecasting*, 34, 477–496.

Baran, S. and Nemoda, D. (2016). Censored and shifted gamma distribution based EMOS model for probabilistic quantitative precipitation forecasting. *Environmetrics*, 27, 280–292.

Bauer, P., Thorpe, A. and Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, 525, 47–55.

Becker, R. and Behrens, K. (2012). Quality assessment of heterogeneous surface radiation network data. *Advances in Science and Research*, 8, 93–97.

Bellier, J., Zin, I. and Bontron, G. (2018). Generating coherent ensemble forecasts after hydrological postprocessing: Adaptations of ECC-based methods. *Water Resources Research*, 54, 5741–5762.

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 (Methodological)*, 57, 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, 988–1002.

Bougeault, P., Toth, Z., Bishop, C., Brown, B., Burridge, D., Chen, D. H., Ebert, B., Fuentes, M., Hamill, T. M., Mylne, K., Nicolau, J., Paccagnella, T., Park, Y.-Y., Parsons, D., Raoult, B., Schuster, D., Dias, P. S., Swinbank, R., Takeuchi, Y., Tennant, W., Wilson, L. and Worley, S. (2010). The THORPEX interactive grand global ensemble. *Bulletin of the American Meteorological Society*, 91, 1059–1072.

Bozonnat, C. and Schlosser, C. (2014). Characterization of the solar power resource in europe and assessing benefits of co-location with wind power installations.

Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and regression trees*. Wadsworth International Group, Belmont.

Bremnes, J. B. (2019a). Constrained quantile regression splines for ensemble postprocessing. *Monthly Weather Review*, 147, 1769–1780.

Bremnes, J. B. (2019b). Ensemble postprocessing using quantile function regression based on neural networks and Bernstein polynomials. *Monthly Weather Review*, 148, 403–414.

Buizza, R. (2018a). Chapter 2 - ensemble forecasting and the need for calibration. In *Statistical Postprocessing of Ensemble Forecasts* (S. Vannitsem, D. S. Wilks and J. W. Messner, eds.). Elsevier, 15–48.

Buizza, R. (2018b). Introduction to the special issue on "25 years of ensemble forecasting". *Quarterly Journal of the Royal Meteorological Society*, 145, 1–11.

Buizza, R., Houtekamer, P. L., Pellerin, G., Toth, Z., Zhu, Y. and Wei, M. (2005). A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Monthly Weather Review*, 133, 1076–1097.

Buizza, R., Leutbecher, M., Isaksen, L. and Haseler, J. (2010). Combined use of eda- and sv-based perturbations in the eps. URL `https://www.ecmwf.int/node/17467`.

Buizza, R., Milleer, M. and Palmer, T. N. (1999). Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 125, 2887–2908.

Calanca, P., Bolius, D., Weigel, A. P. and Liniger, M. A. (2010). Application of long-range weather forecasts to agricultural decision problems in europe. *The Journal of Agricultural Science*, 149, 15–22.

Chen, T. and Guestrin, C. (2016). XGBoost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 785–794.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y. and Li, Y. (2019). *xgboost: Extreme Gradient Boosting*. R package version 0.90.0.1 `https://CRAN.R-project.org/package=xgboost` [Accessed on 13 March 2020].

Cloke, H. and Pappenberger, F. (2009). Ensemble flood forecasting: A review. *Journal of Hydrology*, 375, 613–626.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39, 1–22.

Diebold, F. and Mariano, R. (1995). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*, 13, 253–63.

Duan, Q., Ajami, N. K., Gao, X. and Sorooshian, S. (2007). Multi-model ensemble hydrologic prediction using Bayesian model averaging. *Advances in Water Resources*, 30, 1371–1386.

DWD Climate Data Center (2020). Recent 10-minute station observations of solar incoming radiation, longwave downward radiation and sunshine duration for Germany. `https://opendata.dwd.de/climate_environment/CDC/observations_germany/climate/10_minutes/solar/recent/DESCRIPTION_obsgermany_climate_10min_solar_recent_en.pdf` [Accessed on 15 January 2021].

ECMWF Directorate (2012). Describing ECMWF's forecasts and forecasting system. *ECMWF Newsletter* No. 133 – Autumn 2012.

Ehret, U. and Zehe, E. (2011). Series distance – an intuitive metric to quantify hydrograph similarity in terms of occurrence, amplitude and timing of hydrological events. *Hydrology and Earth System Sciences*, 15, 877–896.

Epstein, E. S. (1969). Stochastic dynamic prediction. *Tellus*, 21, 739–759.

Ewen, J. (2011). Hydrograph matching method for measuring model performance. *Journal of Hydrology*, 408, 178–187.

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, 190–202.

Fraunhofer Institute for Solar Energy Systems (2021). Recent Facts about Photovoltaics in Germany. Technical report. `www.ise.fraunhofer.de/en/publications/studies/recent-facts-about-pv-in-germany.html` [Accessed on 13 October 2021].

Friederichs, P. and Hense, A. (2007). Statistical downscaling of extreme precipitation events using censored quantile regression. *Monthly Weather Review*, 135, 2365–2378.

Friedman, J. H. (2000). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29, 1189–1232.

Gascón, E., Lavers, D., Hamill, T. M., Richardson, D. S., Bouallègue, Z. B., Leutbecher, M. and Pappenberger, F. (2019). Statistical postprocessing of dual-resolution ensemble precipitation forecasts across europe. *Quarterly Journal of the Royal Meteorological Society*, 145, 3218–3235.

Glahn, H. R. and Lowry, D. A. (1972). The use of model output statistics (MOS) in objective weather forecasting. *Journal of Applied Meteorology*, 11, 1203–1211.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106, 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, 243–268.

Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1, 125–151.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102, 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, 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, 411–422.

Gneiting, T. and Ranjan, R. (2013). Combining predictive distributions. *Electronic Journal of Statistics*, 7, 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, 310–323.

Goodfellow, I., Bengio, Y. and Courville, A. (2016). *Deep Learning*. MIT Press.

Gottwalt, S., Gärttner, J., Schmeck, H. and Weinhardt, C. (2016). Modeling and valuation of residential demand flexibility for renewable energy integration. *IEEE Transactions on Smart Grid*, 8, 2565–2574.

Hagedorn, R. (2017). Slowly but surely: Observing and supporting the growing use of ensemble products. `https://www.ecmwf.int/node/17625` [Accessed on 16 October 2021].

Haiden, T., Forbes, R., Ahlgrimm, M. and Bozzo, A. (2015). The skill of ecmwf cloudiness forecasts 14–19.

Haiden, T., Janousek, M., Bidlot, J., Buizza, R., Ferranti, L., Prates, F. and Vitart, F. (2018). *Evaluation of ECMWF forecasts, including the 2018 upgrade*. European Centre for Medium Range Weather Forecasts.

Hamill, T. M., Bates, G. T., Whitaker, J. S., Murray, D. R., Fiorino, M., Galarneau, T. J., Zhu, Y. and Lapenta, W. (2013). NOAA's second-generation global medium-range ensemble reforecast dataset. *Bulletin of the American Meteorological Society*, 94, 1553–1565.

Hamill, T. M. and Scheuerer, M. (2018). Probabilistic precipitation forecast postprocessing using quantile mapping and rank-weighted best-member dressing. *Monthly Weather Review*, 146, 4079–4098.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer New York.

Haupt, S. E., Casado, M. G., Davidson, M., Dobschinski, J., Du, P., Lange, M., Miller, T., Mohrlen, C., Motley, A. and Pestana, R. (2019). The use of probabilistic forecasts: Applying them in theory and practice. *IEEE Power and Energy Magazine*, 17, 46–57.

Haupt, S. E., McCandless, T. C., Dettling, S., Alessandrini, S., Lee, J. A., Linden, S., Petzke, W., Brummet, T., Nguyen, N. and Kosović, B. (2020). Combining artificial intelligence with physics-based methods for probabilistic renewable energy forecasting. *Energies*, 13, 1979.

Hemri, S., Fundel, F. and Zappa, M. (2013). Simultaneous calibration of ensemble river flow predictions over an entire range of lead times. *Water Resources Research*, 49, 6744–6755.

Hemri, S., Haiden, T. and Pappenberger, F. (2016). Discrete postprocessing of total cloud cover ensemble forecasts. *Monthly Weather Review*, 144, 2565–2577.

Hemri, S. and Klein, B. (2017). Analog-based postprocessing of navigation-related hydrological ensemble forecasts. *Water Resources Research*, 53, 9059–9077.

Hemri, S., Lisniak, D. and Klein, B. (2015). Multivariate postprocessing techniques for probabilistic hydrological forecasting. *Water Resources Research*, 51, 7436–7451.

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, 9197–9205.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9, 1735–1780.

Hong, T. and Fan, S. (2016). Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting*, 32, 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.

Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques*. Springer New York.

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., Krüger, F. and Lerch, S. (2019). Evaluating probabilistic forecasts with scoringRules. *Journal of Statistical Software*, 90, 1–37.

Køltzow, M., Casati, B., Bazile, E., Haiden, T. and Valkonen, T. (2019). An NWP model intercomparison of surface weather parameters in the european arctic during the year of polar prediction special observing period northern hemisphere 1. *Weather and Forecasting*, 34, 959–983.

Krzysztofowicz, R. (1999). Bayesian theory of probabilistic forecasting via deterministic hydrologic model. *Water Resources Research*, 35, 2739–2750.

Krzysztofowicz, R. (2001). The case for probabilistic forecasting in hydrology. *Journal of Hydrology*, 249, 2–9.

Kull, M., Perello Nieto, M., Kängsepp, M., Silva Filho, T., Song, H. and Flach, P. (2019). Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox and R. Garnett, eds.), vol. 32. Curran Associates, Inc. `arXiv:1910.12656v1`.

Lang, M. N., Lerch, S., Mayr, G. J., Simon, T., Stauffer, R. and Zeileis, A. (2020). Remember the past: a comparison of time-adaptive training schemes for non-homogeneous regression. *Nonlinear Processes in Geophysics*, 27, 23–34.

Lang, S., Hólm, E., Bonavita, M. and Tremolet, Y. (2019). A 50-member ensemble of data assimilations.

Lauret, P., David, M. and Pinson, P. (2019). Verification of solar irradiance probabilistic forecasts. *Solar Energy*, 194, 254–271.

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, G. and Scott, C. (2012). EM algorithms for multivariate gaussian mixture models with truncated and censored data. *Computational Statistics & Data Analysis*, 56, 2816–2829.

Lerch, S. and Baran, S. (2016). Similarity-based semilocal estimation of post-processing models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66, 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, 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, 21206.

Leutbecher, M., Lock, S.-J., Ollinaho, P., Lang, S. T. K., Balsamo, G., Bechtold, P., Bonavita, M., Christensen, H. M., Diamantakis, M., Dutra, E., English, S., Fisher, M., Forbes, R. M., Goddard, J., Haiden, T., Hogan, R. J., Juricke, S., Lawrence, H., MacLeod, D., Magnusson, L., Malardel, S., Massart, S.,

Sandu, I., Smolarkiewicz, P. K., Subramanian, A., Vitart, F., Wedi, N. and Weisheimer, A. (2017). Stochastic representations of model uncertainties at ECMWF: state of the art and future vision. *Quarterly Journal of the Royal Meteorological Society*, 143, 2315–2339.

Leutbecher, M. and Palmer, T. (2008). Ensemble forecasting. *Journal of Computational Physics*, 227, 3515–3539.

Lindström, G., Johansson, B., Persson, M., Gardelin, M. and Bergström, S. (1997). Development and test of the distributed HBV-96 hydrological model. *Journal of Hydrology*, 201, 272–288.

Mallet, V. and Sportisse, B. (2006). Ensemble-based air quality forecasts: A multimodel approach applied to ozone. *Journal of Geophysical Research: Atmospheres*, 111, paper D18302.

Matuszko, D. (2011). Influence of the extent and genera of cloud cover on solar radiation intensity. *International Journal of Climatology*, 32, 2403–2414.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42, 109–127.

McEvoy, A. J., Markvart, T. and Luis, C. (2012). *Practical handbook of photovoltaics: fundamentals and applications*. Elsevier, Academic Press.

McGovern, A., Elmore, K. L., Gagne, D. J., Haupt, S. E., Karstens, C. D., Lagerquist, R., Smith, T. and Williams, J. K. (2017). Using artificial intelligence to improve real-time decision-making for high-impact weather. *Bulletin of the American Meteorological Society*, 98, 2073–2090.

McLachlan, G. J. and Krishnan, T. (1997). *The EM algorithm and Extensions*. Wiley, New York.

Messner, J. W., Mayr, G. J., Wilks, D. S. and Zeileis, A. (2014). Extending extended logistic regression: Extended versus separate versus ordered versus censored. *Monthly Weather Review*, 142, 3003–3014.

Messner, J. W., Mayr, G. J. and Zeileis, A. (2017). Nonhomogeneous boosting for predictor selection in ensemble postprocessing. *Monthly Weather Review*, 145, 137–147.

Mishra, A. K. (2018). Investigating changes in cloud cover using the long-term record of precipitation extremes. *Meteorological Applications*, 26, 108–116.

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, 73–119.

Montani, A., Cesari, D., Marsigli, C. and Paccagnella, T. (2011). Seven years of activity in the field of mesoscale ensemble forecasting by the COSMO-LEPS system: main achievements and open challenges. *Tellus A: Dynamic Meteorology and Oceanography*, 63, 605–624.

Mureau, R., Molteni, F. and Palmer, T. N. (1993). Ensemble prediction using dynamically conditioned perturbations. *Quarterly Journal of the Royal Meteorological Society*, 119, 299–323.

Murphy, A. H. (1973). Hedging and Skill Scores for Probability Forecasts. *Journal of Applied Meteorology*, 12, 215–223.

Palmer, T. N. (2002). The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Quarterly Journal of the Royal Meteorological Society*, 128, 747–774.

Park, Y.-Y., Buizza, R. and Leutbecher, M. (2008). TIGGE: Preliminary results on comparing and combining ensembles. *Quarterly Journal of the Royal Meteorological Society*, 134, 2029–2050.

Phipps, K., Lerch, S., Andersson, M., Mikut, R., Hagenmeyer, V. and Ludwig, N. (2020). Evaluating ensemble post-processing for wind power forecasts. Preprint, available at `http://arxiv.org/abs/2009.14127`.

Pinson, P. and Girard, R. (2012). Evaluating the quality of scenarios of short-term wind power generation. *Applied Energy*, 96, 12–20.

Pinson, P., Madsen, H., Nielsen, H. A., Papaefthymiou, G. and Klöckl, B. (2009). From probabilistic forecasts to statistical scenarios of short-term wind power production. *Wind Energy*, 12, 51–62.

Pinson, P. and Messner, J. W. (2018). Application of postprocessing for renewable energy. In *Statistical Postprocessing of Ensemble Forecasts* (S. Vannitsem, D. S. Wilks and J. W. Messner, eds.). Elsevier, 241–266.

Politis, D. N. and Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical association*, 89, 1303–1313.

Poulos, H., Chernoff, B., Fuller, P. and Butman, D. (2012). Ensemble forecasting of potential habitat for three invasive fishes. *Aquatic Invasions*, 7, 59–72.

Raftery, A. E., Gneiting, T., Balabdaoui, F. and Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133, 1155–1174.

Rasp, S. and Lerch, S. (2018). Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, 146, 3885–3900.

Ravazzolo, F. and Vahey, S. P. (2014). Forecast densities for economic aggregates from disaggregate ensembles. *Studies in Nonlinear Dynamics & Econometrics*, 18, 367–381.

Reinert, D., Prill, F., Frank, H., Denhard, M., Baldauf, M., Schraff, C., Gebhardt, C., Marsigli, C. and Zängl, G. (2021). DWD Database Reference for the Global and Regional ICON and ICON-EPS Forecasting System. Version 2.1.7. Deutscher Wetterdienst, Offenbach am Main.

Ruiz, J. J. and Saulo, C. (2011). How sensitive are probabilistic precipitation forecasts to the choice of calibration algorithms and the ensemble generation method? part i: sensitivity to calibration methods. *Meteorological Applications*, 19, 302–313.

Ruth, D. P., Glahn, B., Dagostaro, V. and Gilbert, K. (2009). The performance of MOS in the digital age. *Weather and Forecasting*, 24, 504–519.

Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26, 43–49.

Schaake, J. C., Hamill, T. M., Buizza, R. and Clark, M. (2007). HEPEX: The hydrological ensemble prediction experiment. *Bulletin of the American Meteorological Society*, 88, 1541–1548.

Schefzik, R., Thorarinsdottir, T. L. and Gneiting, T. (2013). Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science*, 28, 616–640.

Scheuerer, M. (2013). Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quarterly Journal of the Royal Meteorological Society*, 140, 1086–1096.

Scheuerer, M. and Hamill, T. M. (2015). Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Monthly Weather Review*, 143, 4578–4596.

Schmeits, M. J. and Kok, K. J. (2010). A comparison between raw ensemble output, (modified) Bayesian model averaging, and extended logistic regression using ECMWF ensemble precipitation reforecasts. *Monthly Weather Review*, 138, 4199–4211.

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.

Seibert, S. P., Ehret, U. and Zehe, E. (2016). Disentangling timing and amplitude errors in streamflow simulations. *Hydrology and Earth System Sciences*, 20, 3745–3763.

Sloughter, J. M., Gneiting, T. and Raftery, A. E. (2010). Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *Journal of the American Statistical Association*, 105, 25–35.

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, 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.

Taillardat, M., Fougères, A.-L., Naveau, P. and Mestre, O. (2019). Forest-based and semiparametric methods for the postprocessing of rainfall ensemble forecasting. *Weather and Forecasting*, 34, 617–634.

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, 2375–2393.

Taylor, J. W. and Buizza, R. (2003). Using weather ensemble predictions in electricity demand forecasting. *International Journal of Forecasting*, 19, 57–70.

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, 371–388.

Todini, E. (2008). A model conditional processor to assess predictive uncertainty in flood forecasting. *International Journal of River Basin Management*, 6, 123–137.

Toth, Z. and Kalnay, E. (1993). Ensemble forecasting at NMC: The generation of perturbations. *Bulletin of the American Meteorological Society*, 74, 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.

Vannitsem, S., Bremnes, J. B., Demaeyer, J., Evans, G. R., Flowerdew, J., Hemri, S., Lerch, S., Roberts, N., Theis, S., Atencia, A. et al. (2021). Statistical postprocessing for weather forecasts: Review, challenges, and avenues in a big data world. *Bulletin of the American Meteorological Society*, 102, E681–E699.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer New York.

White, A. A. (2007). The essence of chaos by edward n. lorenz. UCL press. 1993. 227 pp. hardback £16.95. ISBN 1 85728 187 x. *Meteorological Applications*, 1, 289–290.

Wilks, D. S. (2009). Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteorological Applications*, 16, 361–368.

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, 2263–2273.

Wilks, D. S. (2018). Univariate ensemble postprocessing. In *Statistical Postprocessing of Ensemble Forecasts* (S. Vannitsem, D. S. Wilks and J. W. Messner, eds.). Elsevier, 49–89.

Wilks, D. S. (2019). *Statistical Methods in the Atmospheric Sciences*. 4th ed. Elsevier Academic Press.

Williams, R. M., Ferro, C. A. T. and Kwasniok, F. (2013). A comparison of ensemble post-processing methods for extreme events. *Quarterly Journal of the Royal Meteorological Society*, 140, 1112–1120.

World Meteorological Organization (2017). International cloud atlas. manual on the observation of clouds and other meteors. `https://cloudatlas.wmo.int/en/home.html` [Accessed on 13 March 2020].

Wu, X., Zhu, X., Wu, G.-Q. and Ding, W. (2014). Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, 26, 97–107.

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, 022701.

Yang, D. (2020a). Ensemble model output statistics as a probabilistic site-adaptation tool for satellite-derived and reanalysis solar irradiance. *Journal of Renewable and Sustainable Energy*, 12, 016102.

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, 036101.

Ye, Q.-Z. and Chen, S.-S. (2012). The ultimate meteorological question from observational astronomers: how good is the cloud cover forecast? *Monthly Notices of the Royal Astronomical Society*, 428, 3288–3294.

Yuan, X., Wood, E. F. and Ma, Z. (2015). A review on climate-model-based seasonal hydrologic forecasting: physical understanding and system development. *Wiley Interdisciplinary Reviews: Water*, 2, 523–536.

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, 563–579.

Zelikman, E., Zhou, S., Irvin, J., Raterink, C., Sheng, H., Kelly, J., Rajagopal, R., Ng, A. Y. and Gagne, D. (2020). Short-term solar irradiance forecasting using calibrated probabilistic models. NeurIPS Workshop on Tackling Climate Change with Machine Learning, `https://www.climatechange.ai/papers/neurips2020/6` [Accessed on 16 October 2021].

Zhou, X., Zhu, Y., Hou, D., Luo, Y., Peng, J. and Wobus, R. (2017). Performance of the new NCEP global ensemble forecast system in a parallel experiment. *Weather and Forecasting*, 32, 1989–2004.