Short thesis for the degree of doctor of philosophy (PhD)

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

by Mehrez El Ayari

Supervisor: Prof. Dr. Sándor Baran

UNIVERSITY OF DEBRECEN
Doctoral School of Informatics
Debrecen, 2022

Contents

1 Introduction **1**

2 Models and forecast evaluation **5**
2.1 Post-processing of water level forecasts 5
2.2 Post-processing of solar irradiance forecasts 5
2.3 Post-processing of total cloud cover forecasts 7
2.3.1 Multiclass logistic regression 8
2.3.2 Proportional odds logistic regression 9
2.3.3 Multilayer perceptron neural network 9
2.3.4 Random forest models and gradient boosting machines 9
2.3.5 Model formulations 10
2.4 Forecast evaluation 10

3 Data **13**

4 Results and discussion **16**
4.1 Post-processing of water level forecasts 16
4.1.1 BMA Vs. EMOS using RTP 16
4.1.2 Analog-Based Vs. RTP 17
4.1.3 Analog-Based BMA Vs. Analog-Based EMOS 17
4.2 Post-processing of solar irradiance forecasts 18
4.2.1 Results for the AROME-EPS dataset 18
4.2.2 Results for the ICON-EPS dataset 18
4.3 Post-processing of total cloud cover forecasts 20
4.3.1 Post-processing using an extended feature set 21

Bibliography **26**

Publications **27**

1 Introduction

Extreme weather events can lead to humanitarian crises as they have both social and economic impacts on our society. Therefore 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 ensemble forecasts issued differently either in the numerical model or its initial condition. Those forecasts are subject to errors and not reliable in terms of representation of the weather quantity at hand. The necessity 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. Given the ensemble forecasts and their corresponding observations for a given lead time and training period, one can derive distribution based model. 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).

These statistical models or equations used in the post-processing methods, are actually mathematical representations of the relationship between the predictors and the predictands. This has made the opportunity for new applications of these statistical models. In this thesis we present research 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)
- total cloud cover (TCC) of 3330 synoptic (SYNOP) stations around the world.

Ensemble forecasts of these weather quantities are typically highly underdispersive and have systemic bias, where TCC underperform ensemble forecasts of other weather variables.
For water level forecasting, we introduce a doubly truncated Bayesian model averaging (BMA; Raftery et al., 2005) method, which allows for flexible post-processing of possibly multimodel Box-Cox transformed ensemble forecasts of water level. Then we compare its predictive skill to the raw ensemble and the reference truncated normal ensemble model output statistics (EMOS; Gneiting et al., 2005) approach of Hemri and Klein (2017). We considered the following case studies:

- BMA vs. EMOS using rolling training period.
- BMA with rolling vs. analog-based training period selection.
- Analog-based training period selection BMA vs. EMOS.

For ensemble weather predictions of solar irradiance, the proposed models provide probabilistic forecasts in the form of a left-censored logistic (CL0) probability distribution and are evaluated in two case studies covering distinct physical models, geographical regions, temporal resolutions, and types of solar irradiance. The investigated forecasts are:

- AROME-EPS forecasts of the Hungarian Meteorological Service (HMS).
- ICON-EPS forecasts of the Deutsche Wetterdienst (DWD).

As for the total cloud cover prediction, we investigate the performance of post-processing using:

- Multilayer perceptron (MLP) neural networks.
- Gradient boosting machines (GBM).
- Random forest (RF) methods.

Based on the European Centre for Medium-Range Weather Forecasts (ECMWF) global TCC ensemble forecasts we compare the above-listed approaches with:

- The proportional odds logistic regression (POLR).
- The multiclass logistic regression (MLR) models.
- The raw TCC ensemble forecasts

The considered case studies are the following:

- Without incorporating ensemble forecasts of precipitation as additional predictor.
- With incorporating ensemble forecasts of precipitation as additional predictor.

Main results

- Post-processing of water level forecasts:
  - The BMA model outperforms the reference EMOS approach and raw ensemble forecasts considerably using rolling training period.
  - The use the analog-based selection of training periods gives only a small advantage of BMA compared to EMOS.
- Post-processing of solar irradiance forecasts:
  - The proposed CL0-EMOS significantly improves the predictve performance compared both with AROME-EPS and ICON-EPS raw forecasts.

  - The overall level of improvements achieved are comparable to meteorological variables such as precipitation accumulation or total cloud cover.
  - These improvements are of relevance for solar energy forecasting in terms of potential economic benefits and integrating volatile photovoltaic power systems into the electrical grid.

- Post-processing of total cloud cover forecasts:
  - All proposed methods significantly outperform the raw ensemble for all lead times.
  - Seasonally trained models further result in slightly better predictive performance than their non-seasonal counterparts.
  - The use of the mean precipitation accumulation as additional covariate further improves the predictive performance and changes the ranking of the different methods.

2 Models and forecast evaluation

2.1 Post-processing of water level forecasts

The proposed BMA (Baran et al., 2019) predictive PDF is defined as

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

where we use a doubly truncated normal distribution $\mathcal{N}_a^b(\mu, \sigma^2)$, with PDF

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

and $g_{a,b}(x|\, \mu, \sigma) := 0$ 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.

2.2 Post-processing of solar irradiance forecasts

Because of the discrete-continuous nature of solar irradiance, non-negative predictive distributions with positive mass for zero

irradiance are needed. The chosen distribution for solar irradiance forecasts post-processing is the 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

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

or the generalized PDF

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

In the proposed EMOS (Schulz et al., 2021) 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

$$\begin{aligned} \mu &= a_0 + a_1 f_1 + \cdots + a_K f_K + \nu p_0 \\ \sigma &= \exp\left(b_0 + b_1 \log S^2\right), \end{aligned}$$

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

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

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

$$y_t = c_0 + c_1 \sin\left(\frac{2\pi t}{365}\right) + c_2 \cos\left(\frac{2\pi t}{365}\right) + \varepsilon_t \qquad \text{and} \tag{2.1}$$

$$y_t = d_0 + d_1 \sin\left(\frac{2\pi t}{365}\right) + d_2 \cos\left(\frac{2\pi t}{365}\right) + d_3 \sin\left(\frac{4\pi t}{365}\right) + d_4 \cos\left(\frac{4\pi t}{365}\right) + \varepsilon_t, \tag{2.2}$$

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. In the following, we refer to models (2.1) and (2.2) as «*periodic model*» and «*periodic 2 model*», respectively.

One can calculate the h ahead predictions $\widehat{y}$ and $\widehat{f_k}$ of the observation and ensemble members, respectively, using either (2.1) or (2.2), and consider the following modified link function for the location:

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

2.3 Post-processing of total cloud cover forecasts

TCC observations are discrete, therefore post-processing is reduced to classification, whereas forecasts are continuous in the $[0, 1]$ interval. The proposed methods are tested on the operational ECMWF ensemble forecasts, consisting of a high resolution forecast (f_{HRES}), a control forecast (f_{CTRL}) and 50 exchangeable members. The investigated covariates are the HRES forecast

f_{HRES}, the control forecast f_{CTRL}, the mean of the 50 exchangeable ensemble members $\overline{f}_{\text{ENS}}$ (refer to Chapter 3 for data presentation), the ensemble variance

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

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

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

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.

2.3.1 Multiclass logistic regression

In MLR the log-odds of a given class with respect to a fixed reference class 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 total cloud cover with respect to an M-dimensional feature vector $\boldsymbol{x}$ equals

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

$$L_k(\boldsymbol{x}) := \beta_{0k} + \boldsymbol{x}^\top\boldsymbol{\beta}_k, \tag{2.3}$$

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

$$\mathsf{P}\big(Y \leq y_k \mid \boldsymbol{x}\big) = \frac{\mathrm{e}^{\mathcal{L}_k(\boldsymbol{x})}}{1+\mathrm{e}^{\mathcal{L}_k(\boldsymbol{x})}}, \quad \text{with} \quad \mathcal{L}_k(\boldsymbol{x}) := \gamma_{0k} + \boldsymbol{x}^\top \boldsymbol{\gamma}, \quad k = 1, 2, \ldots, n, \tag{2.4}$$

where we assume that $\gamma_{01} < \gamma_{02} < \cdots < \gamma_{0n}$. In this way POLR model (2.4) is more parsimonious than MLR model (2.3), as it has just $n + M$ unknown parameters.

2.3.3 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.

2.3.4 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). 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.

2.3.5 Model formulations

Forecast	Without precipitation		With precipitation	
	Training		Training	
	Non-seasonal	Seasonal	Non-seasonal	Seasonal
Multilayer perceptron neural network	MLP	MLPS	MLP-P	MLPS-P
Gradient boosting machines	GBM	GBMS	GBM-P	GBMS-P
Random forests	RF	RFS	–	–
Proportional odds logistic regression	POLR	POLRS	POLR-P	POLRS-P
Multiclass logistic regresssion	MLR	MLRS	–	–

2.4 Forecast evaluation

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

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

Continuou ranked probability score

For a predictive CDF F and real-valued observation x, the CRPS is defined as

$$\begin{aligned}\text{CRPS}\left(F, x\right) := &\int_{-\infty}^{\infty}\left(F(y) - \mathbb{1}_{\{y \geq x\}}\right)^2 \mathrm{d}y = \int_{-\infty}^{x} F^2(y)\mathrm{d}y \\ &+ \int_{x}^{\infty}\left(1 - F(y)\right)^2 \mathrm{d}y = \mathsf{E}|X - x| - \frac{1}{2}\mathsf{E}|X - X'|,\end{aligned}$$

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.

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

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

Improvements in score

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

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

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

Probability integral transform

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.

Mean absolute error

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).

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).

3 Data

Water level data

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. 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). Ensemble forecasts are initialized at 6 UTC.

Solar irradiance data

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). 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) 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 30 minutes.

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 forecasts are updated four times per day at 00/06/12/18 UTC with lead times of up to 120 hrs (Reinert et al., 2020). 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. The observations are computed based on 10-minute sums of the corresponding variables. The entire dataset used here covers the period 27 December 2018 – 31 December 2020.

Total cloud cover 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) 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. The TCC data set includes data for 3330 SYNOP observation stations. 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

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

After quality control, our additional precipitation data set includes forecast-observation pairs for 2917 SYNOP stations. TCC and precipitation data are available at 2239 of these stations.

Training data selection

We use the most common one which is rolling training period (RTP), where for a certain verification date, we consider forecasts and corresponding observations of the previous n days (100 days in the most of our cases). We also investigate both local and regional training data selection (Thorarinsdottir and Gneiting, 2010).

Analog-based training data selection for water level forecasts

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, while DTW measures similarity by finding the minimum amplitude error between two time series. For each date of verifcation and each of DTW, HMA, and SD, we pick the 100 most similar training dates from the ECMWF ENS mean trajectories

4 Results and discussion

4.1 Post-processing of water level forecasts

4.1.1 BMA Vs. EMOS using RTP

Compared to the raw ensemble, all calibration approaches reduce the mean CRPS and the gap increases when the lead time increases. All presented methods have their maximal skill score at hour 9. For short lead times, the 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. 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. EMOS outperforms significantly the raw ensemble for all lead times except for the first few hours where it underperforms the BMA approaches.
This is fully in accordance with the verification rank histograms of the raw ensemble and PIT histograms of post-processed forecasts:

- 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.

The uniformity of the PIT values can be accepted at a 5% level of significance for only:

- 6 lead times (4,5,6,7,14 and 17 hrs) for naive BMA.
- 9 lead times (5,6,7,14,17,72,75,77 and 79 hrs) for ML BMA.
- 4 lead times (5,6,7 and 9 hrs) for EMOS.

4.1.2 Analog-Based Vs. RTP

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. The uniformity of PIT can be accepted at a 5% level of significance for:

- 25 lead times for BMA SD.
- 21 lead times for BMA HMA.
- 24 lead times BMA DTW.

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.

4.1.3 Analog-Based BMA Vs. Analog-Based EMOS

The skill scores range between 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.
In terms of MAE, the analog-based BMA approaches have lower MAE values than their analog-based EMOS in 77, 82 and 77 cases respectively.
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. 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.

4.2 Post-processing of solar irradiance forecasts

4.2.1 Results for the AROME-EPS dataset

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. The observed improvements are statistically significant. 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%.
Post-processing increases the accuracy of point forecasts as well. The difference in MAE during peak irradiance hours exceeds 20 W/m^2. The bias in the AROME-EPS is mitigated successfully by post-processing. The RMSE of the mean forecasts yields similar conclusions. The proposed post-processing methods effectively corrects the underdispersion and negative bias of the ensemble members. The PIT histograms of EMOS predictive distributions are almost flat, indicating only a small bias for observations between 12 and 24 UTC.

4.2.2 Results for the ICON-EPS dataset

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.
First, we investigate diurnal effects by analyzing the dependence of the mean CRPS of the various forecast models on the time of the observation. At all time points when positive irradiance is

possible, all post-processing methods outperform the raw ensemble forecasts for both direct and diffuse irradiance. More complex models incorporating periodicity exhibit 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, while there is no discernible difference in the skill of the different EMOS models between 6 and 18 UTC. 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 same observations are confirmed when we consider the CRPSS of the various forecast models on the lead time. 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. 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 the least skillful.

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. In terms of mean CRPS there is no significant difference between the various post-processing methods.

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. 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. None of the more complex models,

in particular, consistently outperform the simple MET approach. 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 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, and all EMOS approaches successfully correct the raw ensemble's underdispersion, resulting in nearly perfectly uniform PIT histograms.

4.3 Post-processing of total cloud cover forecasts

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.
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 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.
The proportion of station indicating significant difference in mean CRPS and LogS, the raw ensemble and RFS forecasts are distinctly separated from the other four methods for all lead

times. 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 a slight increase is observed. 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.
The positive effect of post-processing is reflected on the PIT histograms:

- 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 while some bias emerges as the forecast horizon increases.
- 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.

4.3.1 Post-processing using an extended feature set

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, 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 decrease as the lead time increases. 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.

For all lead times, the PIT 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. 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.

Bibliography

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

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

DWD Climate Data Center (2020). Recent 10-minute station observations of solar incoming radiation, longwave downward radiation and sunshine duration for Germany.

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.

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

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., 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.

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.

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

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.

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.

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

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.

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.

Reinert, D., Prill, F., Frank, H., Denhard, M., Baldauf, M., Schraff, C., Gebhardt, C., Marsigli, C. and Zängl, G. (2020). DWD Database Reference for the Global and Regional ICON and ICON-EPSForecasting System. Version 2.1.1. Deutscher Wetterdienst, Offenbach am Main.

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.

Schulz, B., Ayari, M. E., Lerch, S. and Baran, S. (2021). Post-processing numerical weather prediction ensembles for probabilistic solar irradiance forecasting. *Solar Energy*, 220, 1016–1031. URL `https://doi.org/10.1016/j.solener.2021.03.023`.

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.

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.

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

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.

List of Publications

List of Conferences [C]:

C1 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.

C2 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.

C3 Accepted abstract for the German Probability and Statistics Days 2020 (GPSD2020).

C4 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.

DEBRECENI EGYETEM * Egyetemi és Nemzeti Könyvtár *

14 April, 2022