Received: 6 November 2022 | Revised: 14 November 2023 | Accepted: 29 November 2023

DOI: 10.1002/qj.4635

Quarterly Journal of the Royal Meteorological Society RMetS

RESEARCH ARTICLE

A two-step machine-learning approach to statistical post-processing of weather forecasts for power generation

Ágnes Baran | **Sándor Baran**

Faculty of Informatics, University of Debrecen, Debrecen, Hungary

Correspondence
Sándor Baran, Faculty of Informatics, University of Debrecen, Kassai út 26, H-4028 Debrecen, Hungary.
Email: baran.sandor@inf.unideb.hu

Funding information
Hungarian National Research, Development and Innovation Office, Grant/Award Number: K142849

By the end of 2021, the renewable energy share of the global electricity capacity reached 38.3% and the new installations were dominated by wind and solar energy, showing global increases of 12.7% and 18.5% respectively. However, both wind and photovoltaic energy sources are highly volatile, making planning difficult for grid operators; thuss, accurate forecasts of the corresponding weather variables are essential for reliable electricity predictions. The most advanced approach in weather prediction is the ensemble method, which opens the door for probabilistic forecasting. However, ensemble forecasts are often underdispersive and subject to systematic bias. Hence, they require some form of statistical post-processing, where parametric models provide full predictive distributions of the weather variables at hand. We propose a general two-step machine-learning-based approach to calibrating ensemble weather forecasts, where, in the first step, improved point forecasts are generated, which then together with various ensemble statistics serve as input features of the neural network estimating the parameters of the predictive distribution. In two case studies based on 100 m wind speed and global horizontal irradiance forecasts of the operational ensemble prediction system of the Hungarian Meteorological Service, the predictive performance of this novel method is compared with the forecast skill of the raw ensemble and the state-of-the-art parametric approaches. Both case studies confirm that, at least up to 48 hr, statistical post-processing substantially improves the predictive performance of the raw ensemble for all forecast horizons considered. The variants of the proposed two-step method investigated outperform in skill their competitors, and the suggested new approach is well applicable for different weather quantities and for a fair range of predictive distributions.

KEYWORDS

ensemble calibration, ensemble model output statistics, multilayer perceptron, one-dimensional convolutional neural network, solar irradiance, wind speed

1 | INTRODUCTION

The transition from fossil fuels to renewable energy sources is an essential step towards achieving the climate goals and fighting the challenges caused by emission of greenhouse gases and air pollution. By the end of 2021, the renewable energy share of the global electricity capacity reached 38.3%, whereas in the European Union this portion went up to 79.6%. The largest part of 40% of the global renewable capacity still corresponds to traditional hydropower; however, the newly installed capacity is dominated by wind and solar energy, showing global increases of 12.7% and 18.5% respectively (IRENA, 2022). In 2021, wind farms covered 15% of the total electricity demand of the European Union and United Kingdom (Wind Europe, 2022); moreover, according to the independent climate think-tank Ember, during the peak months June and July 2021, for the first time, solar panels generated a tenth of EU electricity (Broadbent, 2021).[1] However, both wind and photovoltaic (PV) energy sources are highly volatile, making planning difficult for grid operators. Hence, according to the scheduling requirements in, for example, Hungary, power plants must indicate in advance in quarter-hour increments how much energy they will produce, which makes accurate short-range prediction of wind and solar power of utmost importance.

Despite the complex relationship between wind speed/solar irradiance and wind/PV energy produced, accurate forecasts of the corresponding weather variables are essential for reliable electricity predictions. Weather forecasts are typically produced with the help of numerical weather prediction models describing the dynamical and physical behaviour of the atmosphere with the help of nonlinear partial differential equations. In order to address uncertainties in the numerical model itself or in the initial conditions, numerical weather prediction models are usually run several times with varying initial conditions or model parametrizations, resulting in a forecast ensemble (Bauer et al., 2015; Buizza, 2018a). In the last 30 years the ensemble method has become a widely used approach all over the world, as the derived probabilistic forecasts allowing the estimation of probability distributions of future weather variables provide an important tool to help forecast-based decision-making (Fundel et al., 2019). Recently, ensemble weather forecasts have also serve as inputs to probabilistic hydrological (Cloke & Pappenberger, 2009) or renewable energy forecasts (Pinson & Messner, 2018).

However, despite the weather centres continuously improving their operational ensemble prediction systems (EPSs), ensemble forecasts still keep suffering from systematic errors, such as lack of calibration or bias (Buizza et al., 2005), and thus require some form of statistical post-processing (Buizza, 2018b). Various approaches to statistical post-processing of a wide variety of weather variables have been proposed over recent years; for a detailed overview of the state-of-the-art techniques, see, for example, Wilks (2018) or Vannitsem et al. (2021).

This article builds on parametric post-processing methods providing full predictive distributions of the weather variables at hand. A popular choice is the non-homogeneous regression or ensemble model output statistics (EMOS; Gneiting et al., 2005), which specifies the predictive distribution by a single parametric law with parameters connected to the ensemble members via appropriate link functions. EMOS models for different weather quantities mainly differ in the parametric family describing the predictive distribution; however, the link functions connecting the distribution parameters to the ensemble forecast might also differ. This computationally efficient method shows excellent performance for a large variety of EPSs, forecast domains and weather variables (see e.g., Wilks, 2019, Sec. 8.3.2); moreover, the majority of the existing EMOS models are implemented in the ensembleMOS R package (Yuen et al., 2018).

Here, we focus on quantities important in wind and solar energy production and investigate statistical post-processing of ensemble forecasts of wind speed measured at hub height (100 m) and global horizontal irradiance (GHI). The case studies presented are based on 11-member short-term ensemble forecasts of these variables produced by the Applications of Research to Operations at Mesoscale EPS (AROME-EPS; Jávorné Radnóczi et al., 2020) of the Hungarian Meteorological Service (HMS). Wind speed calls for EMOS predictive distributions with a non-negative support and positive skew, such as the truncated normal (TN; Thorarinsdottir & Gneiting, 2010), the log-normal (LN; Baran & Lerch, 2015), or the truncated generalized extreme value (TGEV; Baran et al., 2021), whereas the discrete-continuous nature of solar irradiance, similar to precipitation accumulation, requires non-negative predictive distributions assigning positive mass to the event of zero irradiance. The simplest method is to left censor a suitable distribution at zero, an approach that is followed by Scheuerer (2014) and Baran and Nemoda (2016) in precipitation modelling, and recently by Schultz et al. (2021) proposing censored logistic (CL0) and censored normal (CN0) EMOS models for calibrating solar irradiance ensemble forecasts.

To improve the flexibility of parametric post-processing methods, Rasp and Lerch (2018) and Ghazvinian et al. (2021, 2022) apply machine-learning-based techniques for estimation of the unknown parameters of the predictive distributions for temperature and precipitation accumulation respectively. A modified version of this approach appears in Baran and Baran (2021), where

a novel model using a TN predictive distribution for calibrating hub-height short-term AROME-EPS wind-speed ensemble forecasts with parameters connected to the ensemble members via a multilayer perceptron (MLP) neural network (MLP; see e.g., Goodfellow et al., 2016, Chapter 6) significantly outperforms the state-of-the-art EMOS methods. In contrast to Rasp and Lerch (2018) and Ghazvinian et al. (2021, 2022), the proposed approach relies on the same training data as the corresponding EMOS models; that is, neither additional input variables, nor large training datasets are required. For a recent review and systematic comparison of the various machine-learning-based methods in ensemble calibration, we refer the reader to Schultz and Lerch (2022).

In this article, we first extend the technique of Baran and Baran (2021) to CN0 and LN predictive distributions, where the censored Gaussian law is applied for post-processing of solar irradiance, whereas with the LN model wind speed ensemble forecasts are calibrated. Further, we suggest a new general two-step method for estimating the unknown parameters of the TN, LN, and CN0 predictive distributions. Using the same training data as before, with the help of auxiliary neural networks we first provide improved predictions of the weather variables investigated. The required distribution parameters are then obtained in the second step, where the input features of the corresponding neural networks are extended with the previously obtained corrected point forecasts. The predictive performances of the proposed one- and two-step machine-learning-based TN, LN, and CN0 models are compared with those of the corresponding EMOS approaches with matching predictive distributions and with the raw AROME-EPS 100 m wind speed and GHI ensemble forecasts respectively.

2 | DATA

As mentioned, in the case studies of Section 4 we consider two different weather variables used in renewable energy production: 100 m wind speed and GHI. Both datasets investigated contain ensemble forecasts of the AROME-EPS system of the HMS together with the corresponding validating observations. The wind-speed dataset is an extension of forecast–observation pairs investigated in Baran and Baran (2021), whereas part of the solar irradiance data has already been studied in Schultz et al. (2021).

The 11-member AROME-EPS, consisting of a control run and 10 ensemble members obtained from perturbed initial conditions, covers the Transcarpathian Basin with a horizontal resolution of 2.5 km. For both weather quantities studied, ensemble forecasts initialized at 0000 UTC with a forecast horizon of 48 hr and a temporal resolution of 15 min at the nearest grid point to the observation sites are available for the period between May 7, 2020, and July 31, 2021.

Validating observations of 100 m wind speed ($\mathrm{m \cdot s^{-1}}$) with a 15 min temporal resolution are given for three wind farms in the northwestern part of Hungary (Ács, Jánossomorja, and Pápakovácsi) for the period January 1, 2020, to July 5, 2021. The equal temporal resolution of forecasts and observations results in 192 lead times per run. Though the AROME-EPS forecast dataset for these three locations is complete, this is far from the case with data provided by the wind farms, as around 3% of the observations are missing.

GHI observations ($\mathrm{W \cdot m^{-2}}$) are available from two different sources. For modelling purposes, we consider high-quality observations of the HMS with a temporal resolution of 10 min for seven representative locations in Hungary (Aszód, Budapest, Debrecen, Kecskemét, Pécs, Szeged, and Tápiószele) covering the time interval between April 1, 2020, and June 30, 2021. However, we also have observations for two solar farms (Monor and Nagykőrös) with a 15 min temporal resolution for the period May 1, 2020, to August 18, 2021. In the case study of Section 4, data for these two locations, where around 2% of the observations are missing, are merely used for model verification. Owing to the difference in the time steps of the AROME-EPS forecasts and HMS observations, we consider GHI data with a temporal resolution of 30 min in our analysis, resulting in a total of 96 forecast cases per submission.

3 | POST-PROCESSING METHODS AND FORECAST EVALUATION

As mentioned, we consider parametric post-processing approaches resulting in full predictive distributions of the future value of the weather variable investigated, where parameters of these distributions depend on various functions of the corresponding raw ensemble forecast. According to the optimum score principle of Gneiting and Raftery (2007), model fitting is performed by optimizing the mean value of a proper scoring rule (see Section 3.2) over the forecast cases in the training data consisting of past forecast–observation pairs.

In what follows, let $f_1, f_2, \ldots, f_{11}$ denote the 11-member AROME-EPS forecast for a given location, time point, and forecast horizon, where $f_1 = f_{\mathrm{CTRL}}$ is the control forecast and where $f_2, f_3, \ldots, f_{11}$ correspond to ensemble members $f_{\mathrm{ENS},1}, f_{\mathrm{ENS},2}, \ldots, f_{\mathrm{ENS},10}$ generated using perturbed initial conditions. These latter 10 forecasts are statistically indistinguishable and should be treated as exchangeable. Further, denote by $\overline{f}$ the ensemble mean and by $\overline{f}_{\mathrm{ENS}}$ the mean of the 10 exchangeable members, while S^2 and MD stand

for the ensemble variance and ensemble mean absolute difference, respectively, defined as follows:

$$S^2 := \frac{1}{10}\sum_{k=1}^{11}\left(f_k - \bar{f}\right)^2 \quad \text{and} \quad \text{MD} := \frac{1}{11^2}\sum_{k=1}^{11}\sum_{\ell=1}^{11}|f_k - f_\ell|.$$

3.1 | Ensemble model output statistics

The following short review focuses on EMOS models showing the best forecast skill in the preliminary studies with the AROME-EPS 100 m wind speed (Baran & Baran, 2021) and GHI (Schultz et al., 2021) ensemble forecasts.

3.1.1 | EMOS models for wind speed

The first method considered to calibrate AROME-EPS hub-height wind-speed ensemble forecasts is a TN EMOS model with predictive distribution

$$\mathcal{N}_0(a_0 + a_{\text{CTRL}}^2 f_{\text{CTRL}} + a_{\text{ENS}}^2 \bar{f}_{\text{ENS}}, b_0^2 + b_1^2 \text{MD}), \qquad (1)$$

where $\mathcal{N}_0(\mu, \sigma^2)$ denotes a TN distribution with location μ and scale $\sigma > 0$, left-truncated at zero, having probability density function (PDF)

$$g(x|\mu,\sigma) := \frac{1}{\sigma}\frac{\varphi(x-\mu)/\sigma}{\Phi(\mu/\sigma)} \qquad \text{if } x \geq 0$$

and $g(x|\mu,\sigma) := 0$ otherwise, with φ and Φ respectively being the PDF and the cumulative distribution function (CDF) of a standard normal distribution. As mentioned, model parameters $a_0, a_{\text{CTRL}}, a_{\text{ENS}}, b_0, b_1 \in \mathbb{R}$ are estimated according to the optimum score principle.

As an alternative, we also consider the LN EMOS approach of Baran and Lerch (2015), where the mean m and variance v of the LN predictive distribution are linked to the ensemble members via the expressions

$$m = \alpha_0 + \alpha_{\text{CTRL}}^2 f_{\text{CTRL}} + \alpha_{\text{ENS}}^2 \bar{f}_{\text{ENS}} \quad \text{and} \quad v = \beta_0^2 + \beta_1^2 S^2.$$

Similar to the TN EMOS model in Equation (1), to obtain parameters $\alpha_0, \alpha_{\text{CTRL}}, \alpha_{\text{ENS}}, \beta_0, \beta_1 \in \mathbb{R}$ one has to perform an optimum score estimation based on some verification measure.

3.1.2 | EMOS models for solar irradiance

Consider a logistic distribution with location μ and scale $\sigma > 0$ specified by the CDF

$$G(x|\mu,\sigma) := [1 + \mathrm{e}^{-(x-\mu)/\sigma}]^{-1}, \qquad x \in \mathbb{R}.$$

Then, the CDF of the logistic distribution with parameters μ and σ left censored at zero (CL0) is given by

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

In the CL0 EMOS model of Schultz et al. (2021), the link functions connecting the location parameter μ and the scale parameter σ to the ensemble members are given by

$$\begin{aligned} \mu &= \gamma_0 + \gamma_{\text{CTRL}} f_{\text{CTRL}} + \gamma_{\text{ENS}} \bar{f}_{\text{ENS}} + \nu p_0 \quad \text{and} \\ \sigma &= \exp(\delta_0 + \delta_1 \log S^2), \end{aligned} \qquad (2)$$

where p_0 is the proportion of zero forecasts in the ensemble; that is,

$$p_0 := \frac{1}{11}\sum_{k=1}^{11}\mathbb{I}_{\{f_k=0\}},$$

with $\mathbb{I}_H$ denoting the indicator function of a set H. As before, model parameters $\gamma_0, \gamma_{\text{CTRL}}, \gamma_{\text{ENS}}, \nu, \delta_0, \delta_1 \in \mathbb{R}$ are estimated by optimizing an appropriate verification measure over the training data.

As an alternative predictive distribution, one can also consider a normal law with mean μ and standard deviation σ left censored at zero (CN0) and use again the link functions specified by Equation (2). Note that this model, referred to as CN0 EMOS, has already been mentioned in Schultz et al. (2021), and in the case study of Section 4.2 the CL0 and CN0 EMOS approaches exhibit almost equal predictive performance.

3.2 | Parameter estimation

All calibration methods are based on training data of forecast–observation pairs, and the appropriate temporal and spatial composition of this dataset is an important issue in statistical post-processing. In the case studies of Section 4 we consider rolling training periods, where parameter estimates are obtained using ensemble forecasts and corresponding validating observations from the preceding ℓ_{tp} calendar days. This temporal composition ensuring a clear separation of training and verification data is the standard approach in distribution-based post-processing, and the optimal training period length is usually determined after some preliminary data analysis. For the EMOS models of Section 3.1 we treat the different forecast horizons separately, which is in full accordance with the approaches of Baran and Baran (2021) and Schultz et al. (2021). Once the temporal range of the training data is given, one can choose between two traditional approaches to spatial selection (Thorarinsdottir

& Gneiting, 2010). In the regional (or global) approach, model parameters are estimated using training data of the whole ensemble domain and all locations share the same set of parameters. It requires quite short training periods and allows an extrapolation of the predictive distribution to unobserved locations where ensemble forecasts are available. In contrast, local parameter estimation results in distinct parameter estimates for the different stations obtained using only training data of the given station. To avoid numerical stability problems, local models require much longer training periods (for optimal training period lengths for EMOS modelling of different weather quantities see e.g., Hemri et al., 2014), but local models will usually outperform their regional counterparts if the training data are large enough.

As mentioned, the optimum score estimates of the unknown parameters of the various EMOS models minimize the mean value of a scoring rule over the training data. Scoring rules are loss functions that assign numerical values to pairs of predictive distributions (given in the form of a PDF, CDF, or a discrete sample such as an ensemble forecast) and corresponding validating observations. Our EMOS predictive distributions are based on the continuous ranked probability score (CRPS; Wilks, 2019, Sec. 9.5.1), which is one of the most popular proper scoring rules in environmental sciences assessing simultaneously both calibration and sharpness of the probabilistic forecast. The former refers to statistical consistency between the forecasts and the corresponding observations, whereas the latter refers to the concentration of the predictive distribution. The CRPS corresponding to a predictive CDF F and observation $x \in \mathbb{R}$ is defined as

$$\begin{aligned} \mathrm{CRPS}(F,x) &:= \int_{-\infty}^{\infty} [F(y) - \mathbb{I}_{\{y \geq x\}}]^2 \, dy \\ &= \mathsf{E}|X - x| - \frac{1}{2}\mathsf{E}|X - X'|, \end{aligned} \tag{3}$$

where X and X' are independent random variables with CDF F and finite first moment. The CRPS is a negatively oriented score; that is, smaller values imply better forecast skill, and the second line of Equation (3) implies that it can be expressed in the same units as the observation. Note that the CRPS for all distribution families considered in Section 3.1 can be expressed in closed form— Jordan et al. (2019)—allowing a computationally efficient optimization procedure. Finally, for point forecasts the CRPS reduces to the absolute error; that is, the mean CRPS over the training data is replaced by the mean absolute error (MAE). In this case the mean-squared error (MSE) or the root-mean-squared error (RMSE) are also popular loss functions.

3.3 | Machine-learning-based models

Our first approach, which is a simple MLP model (referred to as MLP-S), extends the machine-learning-based TN model of Baran and Baran (2021) to the censored Gaussian predictive distribution of Section 3.1.2 for calibrating solar irradiance ensemble forecasts, and we also consider a wind-speed model with an LN predictive law. Note that initial tests with machine-learning-based estimation of the parameters of the censored logistic predictive distribution for calibrating solar irradiance forecasts were also performed. However, probably due to the more complex form of the loss function, the model training was far less stable than in the case of the censored Gaussian law.

Furthermore, we introduce a more complex two-step approach based on an extended MLP neural network, which in the following sections will be referred to as MLPex.

3.3.1 | MLP-S model

To estimate the unknown parameters of the predictive distributions, an MLP is trained. It is a feedforward neural network, which in our case consists of an input layer, followed by a single hidden layer, with the last layer being the output layer consisting of as many neurons as there are parameters to be estimated. The outputs of this MLP are the parameters of the predictive distribution, whereas as input we consider features derived from the raw ensemble forecast of the weather variable under study (see Sections 4.1.1 and 4.2.1 for the details). This choice of input variables is in contrast to the approach of Rasp and Lerch (2018), where, in addition to the mean and standard deviation of the ensemble forecasts of the 2 m temperature investigated, station-specific information and the ensemble mean and standard deviation of several other weather quantities are used, or to Ghazvinian et al. (2022) augmenting ensemble statistics for a given location with data of the neighbouring ones and with station-specific information as well. Compared with EMOS modelling, a general MLP allows more general connections between the raw forecast and the parameters of the predictive distribution, which might result in better predictive performance (see e.g., Baran & Baran, 2021).

The weights of the network are determined by minimizing the mean CRPS over the training data. Instead of a very long static training set—for example, Ghazvinian et al. (2021) uses daily forecast–observation pairs of 28 years—we consider the same ℓ_{tp}-day-long rolling training periods as in EMOS modelling; see Section 3.2. However, owing to the much larger number of

unknown weights of the MLP, the training of the network requires more training data than the EMOS approaches of Section 3.1. To avoid the increase of the training period length, the different lead times are pooled and just two networks are trained: one for the 0–24 hr forecasts, and another one for the 24–48 hr forecasts. For the implementation details of this initial network, see Sections 4.1.1 and 4.2.1. Note that a similar pooling of lead times is used in the Atmosphere NETwork approach described, for example, in Demaeyer et al. (2023).

3.3.2 | Extended model

Predictive performance of the individual ensemble forecasts can be very different, sometimes independently of the location and forecast horizon, which highly affects the success of distribution fitting. Typically, a less accurate forecast ensemble results in a wider central prediction interval even in the case of a minor change in the corresponding CRPS value. To correct this deficiency, before training the network for the estimation of the distributional parameters, in our second approach we first try to improve the accuracy of point forecasts. For this aim we introduce two different auxiliary neural networks based on the same input features as the one described in Section 3.3.1, both providing additional corrected point forecasts for each location, time point, and forecast horizon. These improved point forecasts are then used as additional input features to the MLPex network providing the parameters of the predictive distribution. The flowchart of this two-step post-processing approach is given in Figure 1.

One of the newly introduced auxiliary networks is again an MLP having a structure quite similar to the first network, optionally with more hidden layers. However, the output layer contains just a single neuron providing

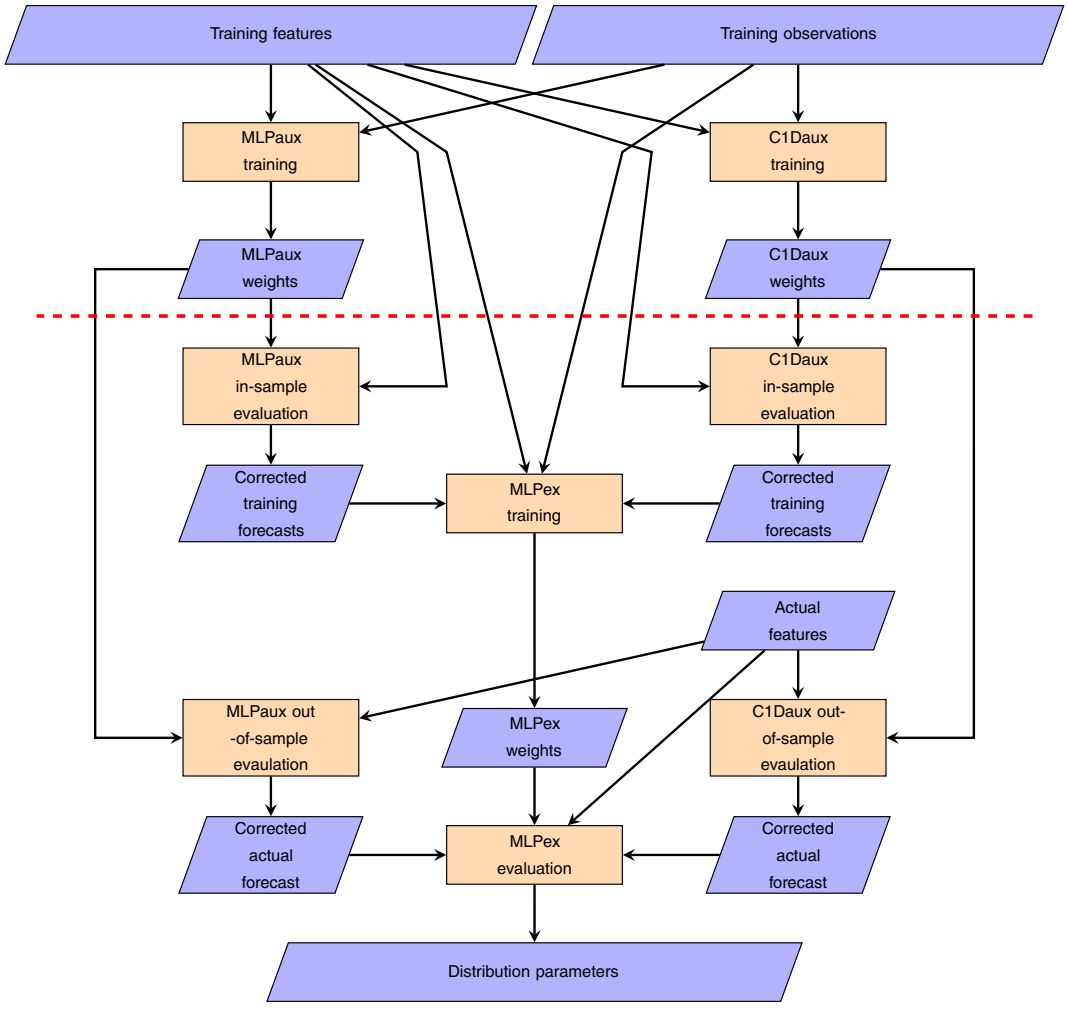

FIGURE 1 Flowchart of the two-step extended multilayer perceptron (MLP) model (MLPex). Model steps are separated by the dashed red line. C1Daux, auxiliary one-dimensional convolutional neural network; MLPaux, auxiliary MLP neural network. [Colour figure can be viewed at wileyonlinelibrary.com]

a point forecast, and the loss function is the MSE or the MAE. This network, referred to as MLPaux, is trained using forecast–observation pairs of the ℓ_{tp}-day-long rolling training period. The trained network is evaluated both for the training features to get the in-sample training forecasts and for the actual features corresponding to the time point of verification resulting in the out-of-sample corrected actual forecast.

The construction of the other auxiliary network is different, as we treat forecasts and observations corresponding to consecutive time points as sequences. Rather than using individual forecast–observation pairs as variables to train the network parameters ignoring time correlations, overlapping short sequences of these pairs are considered, so that the fitted model takes account of short time correlations in the data. For a given location consider separately the $\ell_{\text{tp}}L$-long time series of the input features and the corresponding observations of the given weather variable from the ℓ_{tp}-day long training period having L observations per day measured at L consecutive time points. Then, each of these sequences is split into shorter, overlapping slices of length ℓ_{tw}, considered as time windows. In the case of the features, the first slice is a sequence of vectors consisting of the feature vectors corresponding to the first ℓ_{tw} time steps, and the second slice is obtained by shifting the time window by w, where $w \le \ell_{\text{tw}}$ and w and ℓ_{tw} are hyperparameters of the network. For each sequence of feature vectors, the corresponding sequence of observations is created in a similar way.

For example, consider a three-dimensional input feature vector $(x_1^{(t)}, x_2^{(t)}, x_3^{(t)})^\top$, where index t indicates the time of validity of the corresponding forecasts. Assume that the length of slices is $\ell_{\text{tw}} = 12$ and the shift is $w = 4$. Then, the first two sequences of features are

$$\begin{pmatrix} x_1^{(1)} \\ x_2^{(1)} \\ x_3^{(1)} \end{pmatrix}, \begin{pmatrix} x_1^{(2)} \\ x_2^{(2)} \\ x_3^{(2)} \end{pmatrix}, \ldots, \begin{pmatrix} x_1^{(12)} \\ x_2^{(12)} \\ x_3^{(12)} \end{pmatrix} \text{ and } \begin{pmatrix} x_1^{(5)} \\ x_2^{(5)} \\ x_3^{(5)} \end{pmatrix}, \begin{pmatrix} x_1^{(6)} \\ x_2^{(6)} \\ x_3^{(6)} \end{pmatrix}, \ldots, \begin{pmatrix} x_1^{(16)} \\ x_2^{(16)} \\ x_3^{(16)} \end{pmatrix}.$$

These sequences of length ℓ_{tw} are the inputs of the network, whereas the target vectors are the same slices of the observations.

The first layer of the network is a so-called one-dimensional (1D) convolution layer (see e.g., Kiranyaz et al., 2021) with kernel size κ. It means that a filter window of width κ is sliding along the given input sequence (the height of the window equals the number of input features), and at every position it computes a weighted sum of the underlying values using the weights given in the filter window (weighted moving average). In this way, all κ consecutive feature vectors share the same weight matrix, helping to recognize the temporal behaviour of the process. To enable capturing temporal dependencies in their full complexity, more filter windows of the same size but with different weights are applied simultaneously. The value of κ and the number of filter windows are hyperparameters of the network. The 1D convolutional layer is followed by a pooling layer, where either average or maximum pooling is performed. The aim of a pooling layer is to reduce the size of sequences provided by the 1D convolution layer to get a summarized version of them. This first pooling layer might optionally be followed by further 1D convolution and pooling layers. The last layer before the output layer is a fully connected one, and the number of neurons in the output layer is equal to the length ℓ_{tw} of the target vectors (and to the length of the input feature sequences as well).

To summarize, this network, referred to as C1Daux, provides a sequence-to-sequence estimate; from a sequence of feature vectors of length ℓ_{tw} it computes a sequence of forecasts of the same length. As a loss function the MSE or the MAE is applied, where the computed forecasts and the verifying observations (the target vectors) are compared.

After training this second network with data from the ℓ_{tp}-day long training period, we create an in-sample corrected training prediction time series in the following way. We consider the same $\ell_{\text{tp}}L$-long vector time series of the input features as before and divide it into slices of length ℓ_{tw} again; however, now the slices are not overlapping. The same procedure is repeated with the vector time series corresponding to the L time points of the validation day, resulting in a corrected out-of-sample forecast.

Finally, to estimate the parameters of the predictive distributions, we consider an MLP neural network similar to the one described in Section 3.3.1; however, we extend the set of the input features with the corrected point forecasts obtained using the two auxiliary networks MLPaux and C1Daux. The input features used to train the MLPex network are the same as for the MLP-S extended by the in-sample corrected training forecasts, whereas features derived from the actual forecasts and the two out-of-sample corrected forecasts are applied to obtain the parameters of the actual predictive distribution.

3.4 | Forecast evaluation

The general aim of probabilistic forecasting, as formulated by Gneiting et al. (2007), is to access the maximal sharpness of the predictive distribution subject to calibration.

One of the simplest tools for assessing calibration of ensemble forecasts is the verification rank histogram displaying the ranks of the verifying observations with respect to the corresponding ensemble forecasts (Wilks, 2019, Sec. 9.7.1). For a properly calibrated K-member ensemble,

each of the $K+1$ ranks is equally likely, resulting in a completely flat uniform histogram. The continuous counterpart of the verification rank histogram for probabilistic forecasts specified by predictive distributions is the probability integral transform (PIT) histogram (Wilks, 2019, Sec. 9.5.4). The PIT is defined as the value of the predictive CDF evaluated at the verifying observation; and for a calibrated predictive distribution, PIT values follow a standard uniform law.

However, proper scoring rules allow us to address calibration and sharpness simultaneously. In the case studies of Section 4 the predictive performance of the competing probabilistic forecasts with a given forecast horizon is compared with the help of the mean CRPS over all forecast cases in the verification period. Further, the improvement in terms of the mean CRPS of a probabilistic forecast F with respect to a reference forecast F_{ref} can be quantified using the continuous ranked probability skill score (CRPSS; see e.g., Gneiting & Raftery, 2007). Denoting by $\overline{\text{CRPS}}_F$ and $\overline{\text{CRPS}}_{F_{\text{ref}}}$ the mean CRPS corresponding to forecasts F and F_{ref} respectively, the CRPSS is defined as

$$\text{CRPSS} := 1 - \frac{\overline{\text{CRPS}}_F}{\overline{\text{CRPS}}_{F_{\text{ref}}}}.$$

Note that skill scores are positively oriented; that is, larger CRPSS means better predictive performance compared with the reference forecast.

Calibration and sharpness of a probabilistic forecast can also be investigated with the respective help of the coverage and the average width of the $(1-\alpha)100\%$, $\alpha \in]0,1[$, central prediction intervals. As coverage we consider the proportion of validating observations located between the lower and upper $\alpha/2$ quantiles of the predictive distribution, which for a calibrated forecast should be approximately $(1-\alpha)100\%$ (see e.g., Gneiting & Raftery, 2007). In order to provide a fair comparison with a K-member ensemble forecast, level α should be chosen to match the nominal coverage of $(K-1)/(K+1)100\%$ of the ensemble range (83.33% for the 11-member AROME-EPS).

Finally, point forecasts, such as the median and mean of the raw ensemble and of the calibrated predictive distributions, are evaluated with the use of MAEs and RMSEs, and one should remark that the MAE is optimal for the median, whereas the RMSE is optimal for the mean (Gneiting, 2011).

4 | CASE STUDIES

The efficiency of the simple and extended MLP approaches, proposed in Sections 3.3.1 and 3.3.2 respectively, is tested in two different case studies based on short-term AROME-EPS ensemble forecasts of 100 m wind speed and GHI introduced in Section 2. In both cases, forecast–observation pairs of a whole year from July 1, 2020, to June 30, 2021, are used for model verification, where the forecast skill of the machine-learning-based methods is compared with that of the matching EMOS models and the raw ensemble predictions. As mentioned, in the case of EMOS models each forecast horizon is treated separately, whereas for the MLP-S and MLPEx approaches one network is trained for the 0–24 hr forecasts and another one for the 24–48 hr predictions. For wind speed, this results in 192 sets of EMOS model parameters corresponding to a given forecast initialization time, whereas for solar irradiance one has 96 different sets.

4.1 | Wind speed

Calibration of 100 m wind-speed ensemble forecasts is performed locally using 51-day rolling training periods, which training period length is derived from a detailed preliminary data analysis—for details, see Baran and Baran (2021).

4.1.1 | Implementation details of the neural networks

The input features of the MLP-S network are the control member f_{CTRL}, the mean of the exchangeable members $\bar{f}_{\text{ENS}}$, and the ensemble standard deviation S. The same three features are used to train the C1Daux network, whereas the MLPaux is based on the ensemble mean $\bar{f}$ and standard deviation S. The network MLPex has five input features; namely, the same three as MLP-S, extended with the point forecasts provided by MLPaux and C1Daux.

The main hyperparameters of the different networks are as follows. MLP-S and MLPex have the same architecture: one hidden layer with 28 neurons. Here, the batch size is 1,024, the initial learning rate is 0.01, which is halved in the epochs 8, 28, 48, and 68, and the Adam optimizer is used. In the output layer there are two neurons providing the estimation of the values e^{μ} and e^{σ}, resulting in the location μ and the scale σ of the predictive distribution respectively. Finally, in the input and hidden layers we use the exponential linear unit activation function, whereas in the output layer the linear activation function is applied. MLPaux has two hidden layers, one with five neurons and the other with 15 neurons, the batch size is 1,024, the initial learning rate is 0.01, and this is multiplied by 0.97 in the epochs $3, 4, \ldots, 59$. The network is trained by minimizing the MAE using the Adam optimizer. The C1Daux has a 1D convolutional layer with 24 filters, and the kernel

size $\kappa = 3$. The next layer is a maximum pooling layer with a pool size of 2, which is followed by a flatten layer and a dense layer with 25 neurons. The batch size is 512, and the learning rate, the loss function, and the optimizer are the same as for the MPLaux. Considering the input sequences, $\ell_{\text{tw}} = 16$ and $w = 4$ (see Section 3.3.2). The auxiliary networks use the linear activation function in the output layer and the rectified linear unit function in the other layers.

For all networks, the rolling training data comprising forecast–observation pairs of 51 days preceding the actual date to be modelled are randomly divided into two parts, with 80% of the data used for optimization of the network's parameters and the remaining 20% for the stopping criterion. After each epoch, the loss function is evaluated on this second part of the dataset; and if this value is increasing in 10 subsequent epochs, then the training will terminate.

4.1.2 | Post-processing of 100 m wind-speed ensemble forecasts

Figure 2a shows that, in terms of the mean CRPS, all the post-processing models considered outperform the raw 100 m wind-speed ensemble forecasts for all lead times by a wide margin. A better insight into the differences between the various calibration approaches can be obtained from Figure 2b, which shows the CRPSS values with respect to the TN EMOS model. In general, machine-learning-based methods outperform the corresponding EMOS models for all lead times, and in most cases the MLPex results in a higher skill score than the matching MLP-S. According to Table 1, which provides the overall mean CRPS of post-processed forecasts as a proportion of the mean CRPS of the AROME-EPS, the TN MLPex results in the lowest mean score value, closely followed by the TN MLP-S, LN MLPex, and LN MLP-S.

Figure 3 displays the coverage and the average width of the nominal 83.33% central prediction intervals as functions of the forecast horizon. All post-processed forecasts result in a coverage rather close to the nominal level, whereas the coverage of the raw ensemble is between 51.72% and 70.43% and shows an increasing trend as the lead time increases (see Figure 3a). The mean absolute deviations from the nominal coverage of 83.33% are summarized in Table 2, which provides a possible ranking of the different forecasts. Obviously, the improved coverage of post-processed forecasts is a result of the wider central prediction intervals, as depicted in Figure 3b. All machine-learning approaches provide smooth curves

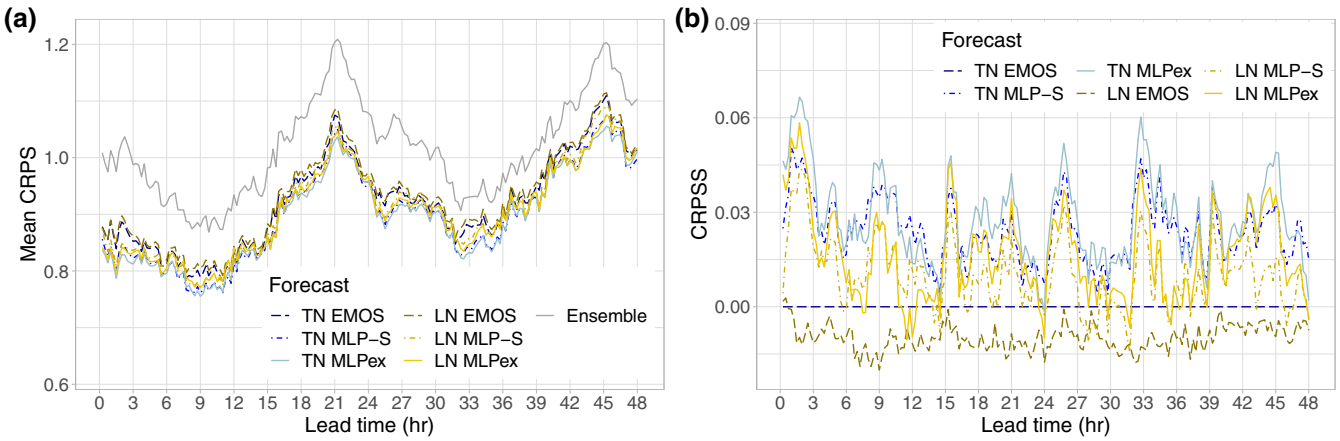

FIGURE 2 (a) Mean continuous ranked probability score (CRPS) of post-processed and raw 100 m wind speed ensemble forecasts and (b) continuous ranked probability skill score (CRPSS) of post-processed forecasts with respect to the TN EMOS model as functions of the lead time. EMOS, ensemble model output statistics; LN, log-normal; MLP, multilayer perceptron; MLP-S, simple MLP model; MLPex, extended MLP model; TN, truncated normal. [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 1 Overall mean continuous ranked probability score (CRPS) of post-processed 100 m wind-speed forecasts as a proportion of the mean CRPS of the Applications of Research to Operations at Mesoscale ensemble prediction system.

TN EMOS	TN MLP-S	TN MLPex	LN EMOS	LN MLP-S	LN MLPex
90.44%	88.30%	87.91%	91.33%	89.55%	88.94%

Abbreviations: EMOS, ensemble model output statistics; LN, log-normal; MLP, multilayer perceptron; MLP-S, simple MLP model; MLPex, extended MLP model; TN, truncated normal.

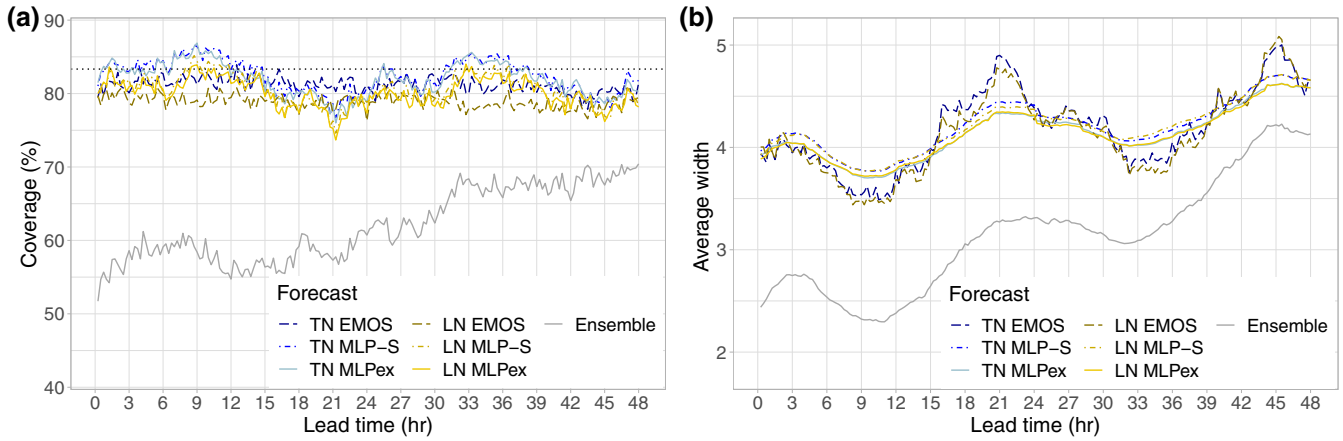

FIGURE 3 (a) Coverage and (b) average width of the nominal 83.33% central prediction intervals of post-processed and raw 100 m wind-speed ensemble forecasts as functions of the lead time. EMOS, ensemble model output statistics; LN, log-normal; MLP, multilayer perceptron; MLP-S, simple MLP model; MLPex, extended MLP model; TN, truncated normal. [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 2 Mean absolute deviation in coverage from the nominal 83.33% level over all lead times.

TN EMOS	TN MLP-S	TN MLPex	LN EMOS	LN MLP-S	LN MLPex	Ensemble
2.18%	1.91%	2.09%	4.44%	2.91%	3.29%	21.30%

Abbreviations: EMOS, ensemble model output statistics; LN, log-normal; MLP, multilayer perceptron; MLP-S, simple MLP model; MLPex, extended MLP model; TN, truncated normal.

of average width, which is a result of using just two different sets of trained networks, whereas each lead time is considered separately in EMOS modelling. From the MLP models, the extended ones result in the sharpest predictive distributions, outperforming the EMOS methods for lead times 15–31 and 41–46 hr.

The better calibration of post-processed forecasts can also be observed in Figure 4, which shows their PIT histograms together with the verification rank histograms of the raw ensemble corresponding to four different lead time intervals. Note that as there are no visible differences between the PIT histograms of the MLPex predictive distributions and the corresponding MLP-S approaches, the latter ones are not shown. The verification rank histograms of the AROME-EPS are symmetric and rather U-shaped. The former indicates the lack of bias in the ensemble forecasts, whereas the latter is a sign of a strongly underdispersive character; however, the dispersion improves with the forecast lead time. The underdispersion is nicely corrected by post-processing, especially when a TN predictive distribution is utilized. For models with an LN predictive distribution, the moment-based α^0_{1234} test (Knüppel, 2015) rejects uniformity at a 5% level of significance for all available lead times, whereas for the TN EMOS, TN MLP-S, and TN MLPex approaches the acceptance rate is 36.5%, 47.4% and 40.6% respectively.

Finally, according to Figure 5, where the difference in MAE of the median forecasts and in RMSE of the mean forecasts from the raw AROME-EPS are depicted as functions of the forecast horizon, post-processing does not really improve the accuracy of point predictions. This is in line with the findings of Baran and Baran (2021) and might be explained by the lack of bias on the ensemble forecasts, meaning no further bias correction is required.

On the basis of the aforementioned results, one can conclude that for both the predictive distributions investigated the machine-learning-based approaches outperform their EMOS counterparts and the extended MLP exhibits slightly better forecast skill than the simple MLP. For the dataset at hand, the best-performing post-processing methods are the novel TN MLPex and the TN MLP-S, resulting in the lowest CRPS, the best coverage, and the most uniformly distributed PIT values.

4.2 | Solar irradiance

In contrast to wind speed, AROME-EPS forecasts of GHI are calibrated regionally, allowing extrapolation of the models investigated to sites excluded from the training. According to the data analysis performed by Schultz et al. (2021), a 31-day rolling training period is applied.

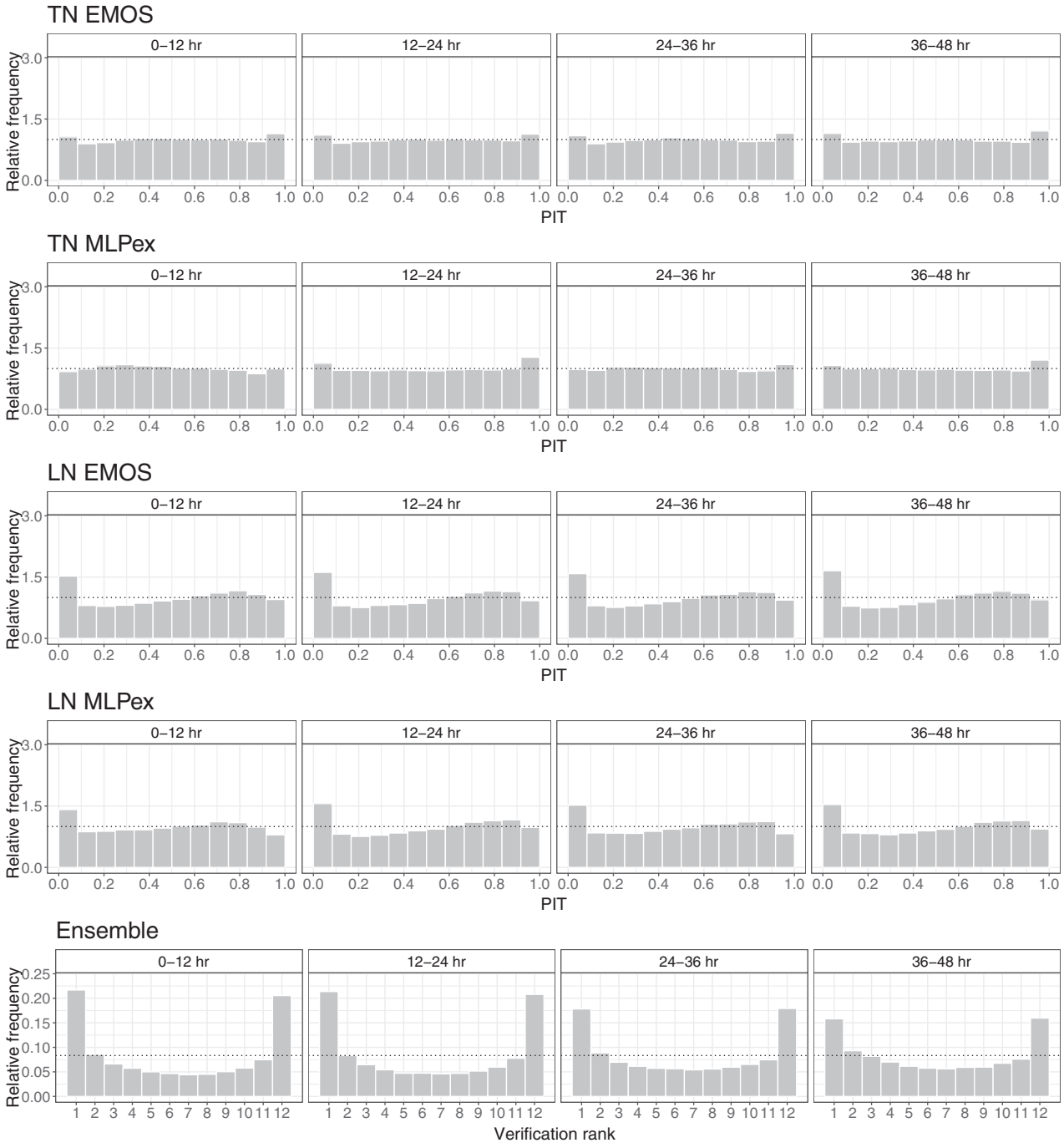

FIGURE 4 Probability integral transform (PIT) histograms of post-processed and verification rank histograms of raw 100 m wind-speed ensemble forecasts for lead times 0–12, 12–24, 24–36, and 36–48 hr. EMOS, ensemble model output statistics; LN, log-normal; MLP, multilayer perceptron; MLP-S, simple MLP model; MLPex, extended MLP model; TN, truncated normal.

4.2.1 | Implementation details of the neural networks

The MLP-S network applied for calibrating solar irradiance ensemble forecasts is trained by five input features, which are the control member f_{CTRL}, the mean of the exchangeable members $\overline{f}_{\mathrm{ENS}}$, the ensemble standard deviation S, an integer between 0 and 47 corresponding to the forecast horizon, and the proportion p_0 of zero forecasts in the ensemble. The MLPaux network is based on the same features as the MLP-S except for p_0, whereas the C1Daux network uses only the mean and the standard deviation of the 11-member ensemble. Finally, the net MLPex works with seven input features; besides the five considered also by MLP-S, it uses the corrected point forecasts provided by the two auxiliary networks.

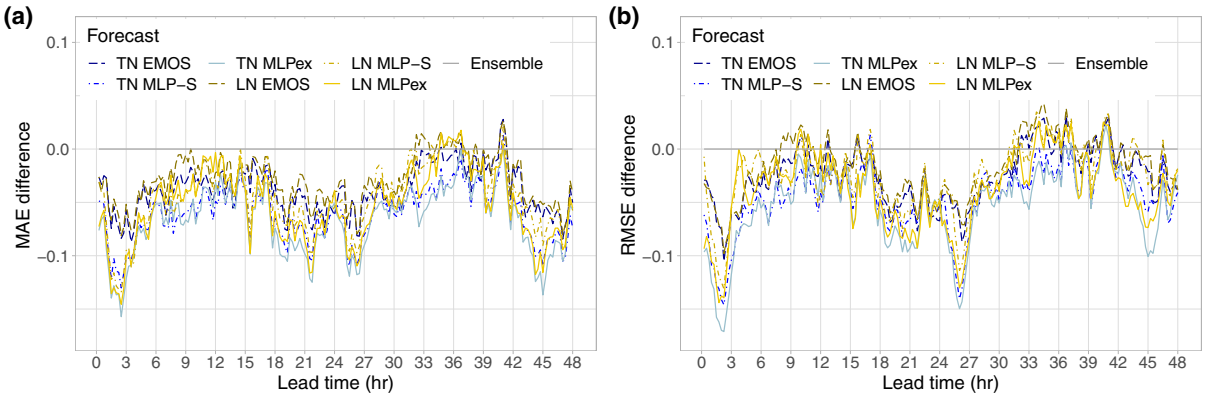

FIGURE 5 Difference in (a) mean absolute error (MAE) of the median forecasts and (b) root-mean-squared error (RMSE) of the mean forecasts from the raw ensemble as functions of the lead time. EMOS, ensemble model output statistics; LN, log-normal; MLP, multilayer perceptron; MLP-S, simple MLP model; MLPex, extended MLP model; TN, truncated normal. [Colour figure can be viewed at wileyonlinelibrary.com]

The structures of the networks are as follows. Both MLP-S and MLPex have only one hidden layer with 35 neurons and an exponential activation function. In the output layer there are two neurons providing the estimation of the values μ^3 and e^σ, and here the activation function is the linear one. MLPaux has one hidden layer with 32 neurons and with a rectified linear unit activation function, followed by a normalization layer, and by the output layer with a linear activation function. The C1Daux has a 1D convolutional layer with 35 filters, and the kernel size $\kappa = 5$. The next layer is an average pooling layer with a pool size of 2, followed by a second 1D convolutional layer with 16 filters and kernel size $\kappa = 2$, a flatten layer, and a dense layer with 30 neurons. Here, the input sequences are of length $\ell_{\text{tw}} = 12$ and the shift is $w = 1$.

Similar to wind speed, 80% of the data of the 31-day rolling training period is used for optimizing the network parameters and the remaining randomly chosen 20% is used for the stopping criterion.

4.2.2 | Post-processing of GHI ensemble forecasts

Figure 6 shows the mean CRPS of post-processed and raw GHI ensemble forecasts and the CRPSS with respect to the CL0 EMOS model suggested by Schultz et al. (2021). Compared with the raw AROME-EPS ensemble, all calibration methods result in a substantial decrease in the mean CRPS between 0300 and 1900 UTC, when positive GHI is likely to be observed. According to Figure 6b, there is just a minor difference in skill between the CL0 and CN0 EMOS models, whereas the two MLP-based approaches result in positive CRPSS in more than 75% of the lead times considered and perform particularly well in the "dark" hours (0000–0300 and 1900–2400 UTC). However, from the point of view of PV energy production, one is interested in the predictive performance of the various forecasts when the observed GHI exceeds some positive threshold. To address this problem, Table 3 summarizes the mean CRPS of post-processed forecasts as a proportion of the mean CRPS of the AROME-EPS for forecast cases with observed irradiance not less than 7.5 $\text{W}\cdot\text{m}^{-2}$, a threshold that had been suggested by forecasters at the HMS. For these cases, the highest skill corresponds to the CN0 MLPex approach, followed by the CN0 MLP-S, which is still more than 2.8% ahead of the best-performing EMOS method.

The improved calibration of post-processed forecasts can be observed in Figure 7a as well, showing the coverage of the nominal 83.33% central prediction intervals. At the hours of peak irradiance (0600–1500 UTC) all calibrated forecasts result in a coverage very close to the nominal value, whereas the coverage of the raw AROME-EPS forecasts hardly exceeds 40%. In general, the machine learning-based approaches outperform the EMOS models, which conclusion is also supported by Table 4 providing the mean absolute deviation in coverage from the nominal 83.33%. Again, as depicted in Figure 7b, the cost of the higher coverage of calibrated forecasts should be paid in the deterioration of the sharpness. From the competing post-processing methods, between 1000 and 1400 UTC the CN0 MLPex results in far the narrowest central prediction intervals, followed by the CN0 MLP-S, whereas outside these hours there is no much difference in sharpness between the various approaches.

Figure 8 showing the verification rank histograms of the raw and the PIT histograms of the calibrated GHI ensemble forecasts for lead times 0–12, 12–24, 24–36, and 36–48 hr also attests to the positive effect of post-processing. The raw AROME-EPS forecasts seem to be negatively biased, tending to underestimate the observed GHI and indeed, the average biases of the

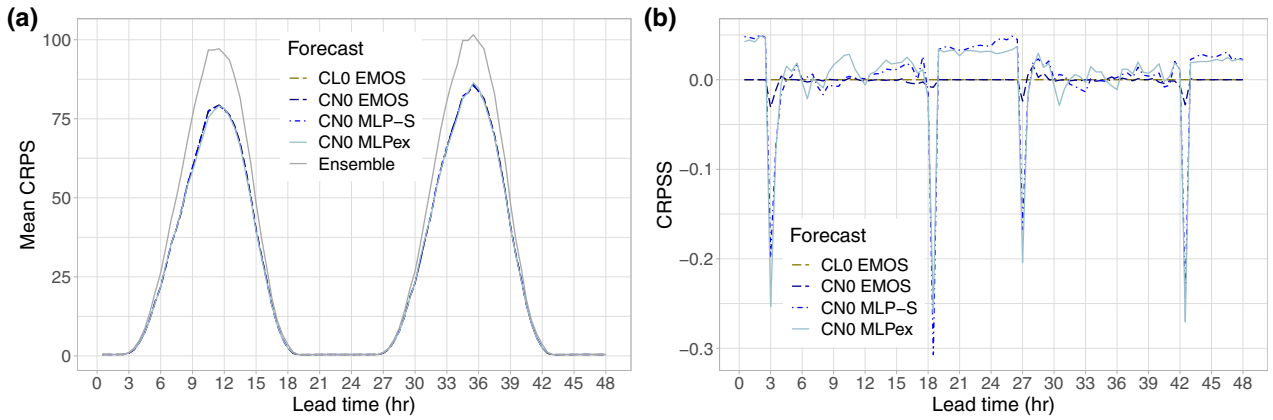

FIGURE 6 (a) Mean continuous ranked probability score (CRPS) of post-processed and raw global horizontal irradiance ensemble forecasts and (b) continuous ranked probability skill score (CRPSS) of post-processed forecasts with respect to the CL0 EMOS model as functions of the lead time. CL0, censored logistic; CN0, censored normal; EMOS, ensemble model output statistics; MLP, multilayer perceptron; MLP-S, simple MLP model; MLPex, extended MLP model. [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 3 Overall mean continuous ranked probability score (CRPS) of post-processed global horizontal irradiance (GHI) forecasts as a proportion of the mean CRPS of the Applications of Research to Operations at Mesoscale ensemble prediction system for observed GHI not less than 7.5 $W \cdot m^{-2}$.

CL0 EMOS	CN0 EMOS	CN0 MLP-S	CN0 MLPex
82.64%	82.67%	79.80%	79.11%

Abbreviations: CL0, censored logistic; CN0, censored normal; EMOS, ensemble model output statistics; MLP, multilayer perceptron; MLP-S, simple MLP model; MLPex, extended MLP model.

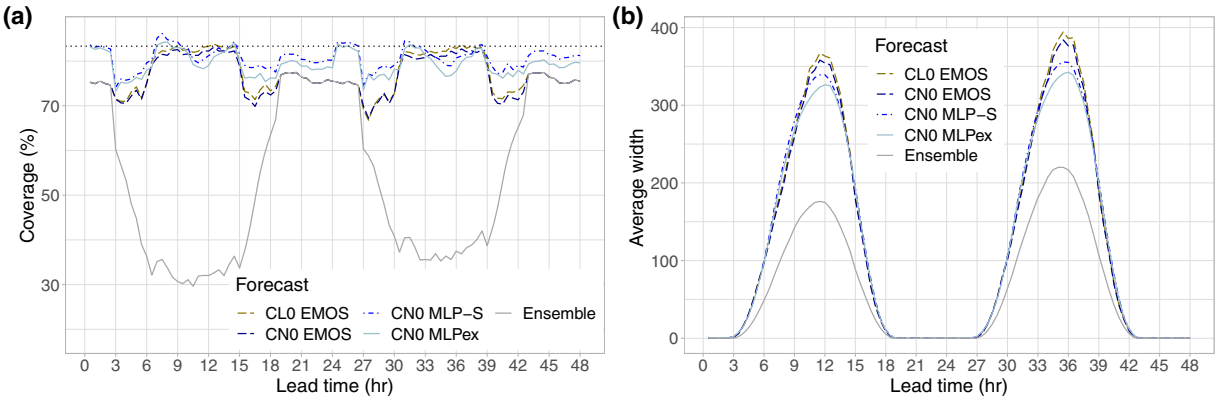

FIGURE 7 (a) Coverage and (b) average width of the nominal 83.33% central prediction intervals of post-processed and raw global horizontal irradiance ensemble forecasts as functions of the lead time. CL0, censored logistic; CN0, censored normal; EMOS, ensemble model output statistics; MLP, multilayer perceptron; MLP-S, simple MLP model; MLPex, extended MLP model. [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 4 Mean absolute deviation in coverage from the nominal 83.33% level over all lead times.

CL0 EMOS	CN0 EMOS	CN0 MLP-S	CN0 MLPex	Ensemble
6.11%	6.64%	2.88%	4.03%	29.53%

Abbreviations: CL0, censored logistic; CN0, censored normal; EMOS, ensemble model output statistics; MLP, multilayer perceptron; MLP-S, simple MLP model; MLPex, extended MLP model.

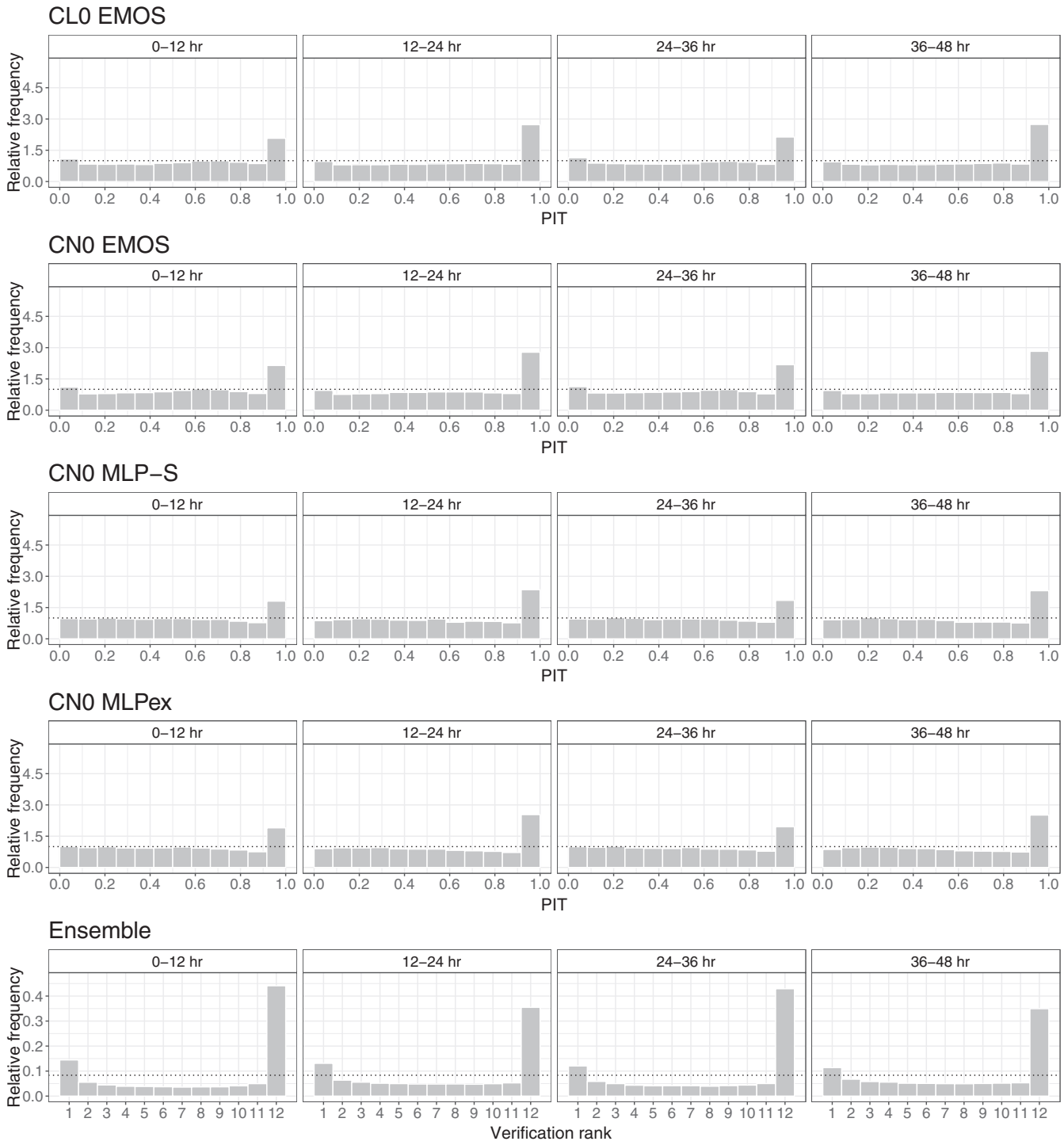

FIGURE 8 Probability integral transform (PIT) histograms of post-processed and verification rank histograms of raw global horizontal irradiance ensemble forecasts for the lead times 0–12, 12–24, 24–36, and 36–48 hr. CL0, censored logistic; CN0, censored normal; EMOS, ensemble model output statistics; MLP, multilayer perceptron; MLP-S, simple MLP model; MLPex, extended MLP model.

ensemble median and the ensemble mean taken over all lead times and forecast cases with observed GHI not less than $7.5\,\mathrm{W{\cdot}m^{-2}}$ are $-16.5\,\mathrm{W{\cdot}m^{-2}}$ and $-20.8\,\mathrm{W{\cdot}m^{-2}}$ respectively. Further, the U-shape of the verification rank histograms indicates strong underdispersion, which is in complete accordance with the low coverage values and sharp prediction intervals observed in Figure 7. All post-processing approaches substantially improve the calibration and reduce the bias; however, compared with the case of wind speed (Section 4.1.2), there are much less lead times, where the α^0_{1234} test accepts uniformity at a 5% level of significance. The highest acceptance rate of 4.2%

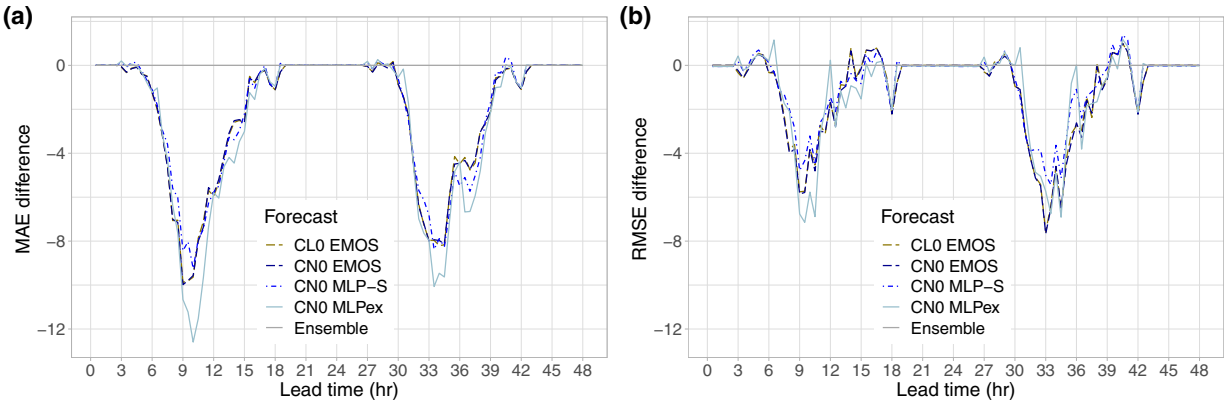

FIGURE 9 Difference in (a) mean absolute error (MAE) of the median forecasts and (b) root-mean-squared error (RMSE) of the mean forecasts from the raw ensemble as functions of the lead time. CL0, censored logistic; CN0, censored normal; EMOS, ensemble model output statistics; MLP, multilayer perceptron; MLP-S, simple MLP model; MLPex, extended MLP model. [Colour figure can be viewed at wileyonlinelibrary.com]

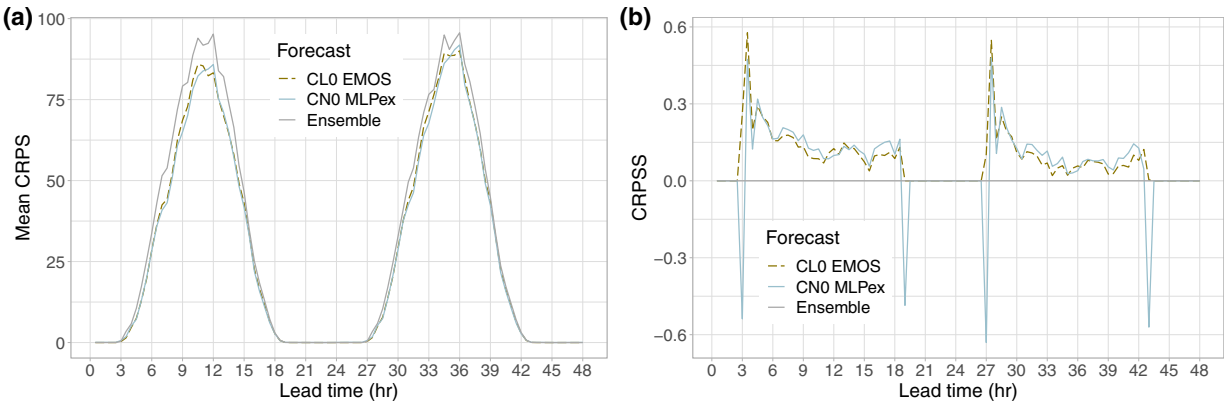

FIGURE 10 (a) Mean continuous ranked probability score (CRPS) of post-processed and raw global horizontal irradiance ensemble forecasts for Monor and Nagykőrös and (b) continuous ranked probability skill score (CRPSS) of post-processed forecasts with respect to the raw ensemble as functions of the lead time. CL0, censored logistic; CN0, censored normal; EMOS, ensemble model output statistics; MLP, multilayer perceptron; MLPex, extended MLP model. [Colour figure can be viewed at wileyonlinelibrary.com]

corresponds to the CN0 MLP-S approach (14, 14.5, 38, 38.5 hr; observations at 1400 and 1430 UTC), followed by the CL0 EMOS (38, 38.5 hr) and the CN0 MLPex (38 hr), whereas for the CN0 EMOS there is no lead time with significantly uniform PIT.

Finally, Figure 9a plots the difference in MAE of the median forecasts of the various post-processing methods from the MAE of the raw ensemble as functions of the forecast horizon. Between 0900 and 1500 UTC all four approaches investigated result in a substantial improvement in accuracy, and the CN0 MLPex method seems to lead to the largest gain in MAE. Fairly similar conclusions can be drawn from Figure 9b showing the difference in RMSE of the median forecasts, and both figures support the presence of a bias in the AROME-EPS that is then corrected by post-processing.

Similar to Section 4.1.2, the results provided here confirm the superiority of the machine-learning-based methods over EMOS modelling. The CN0 MLPex approach results in the lowest mean CRPS, the largest improvement in MAE combined with a fair coverage, and at the hours around 1200 UTC far narrower central prediction intervals than its competitors.

4.2.3 | Model verification for additional locations

All approaches to post-processing GHI ensemble forecasts investigated in Section 4.2.2 are trained regionally using forecast–observations pairs provided by the HMS for seven locations. Hence, one can use the EMOS models and trained neural networks obtained to calibrate AROME-EPS ensemble forecasts for the solar farms in Monor and Nagykőrös, data for which are not used in the training process. The raw and post-processed forecasts for

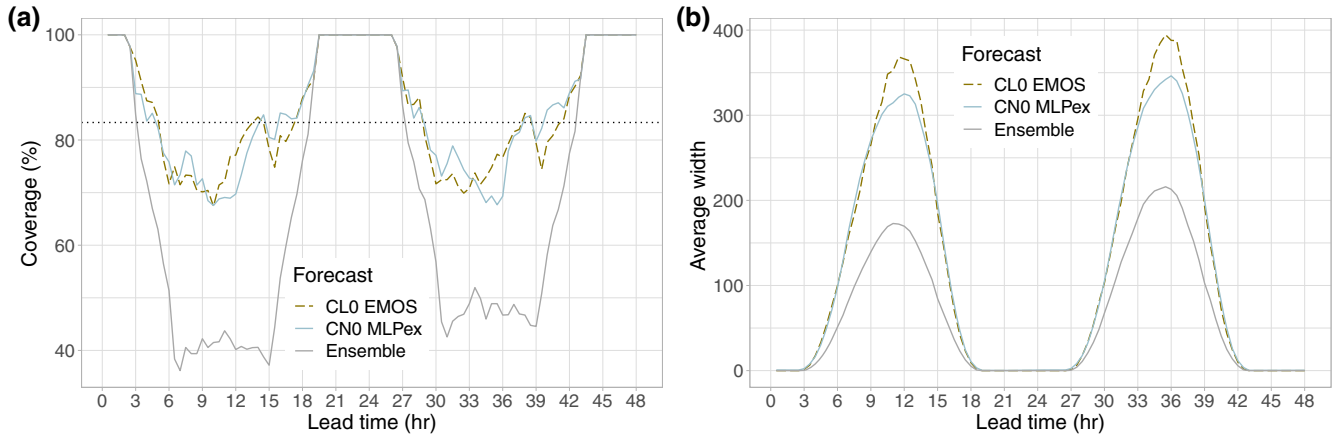

FIGURE 11 (a) Coverage and (b) average width of the nominal 83.33% central prediction intervals of post-processed and raw global horizontal irradiance ensemble forecasts for Monor and Nagykőrös as functions of the lead time. CL0, censored logistic; CN0, censored normal; EMOS, ensemble model output statistics; MLP, multilayer perceptron; MLPex, extended MLP model. [Colour figure can be viewed at wileyonlinelibrary.com]

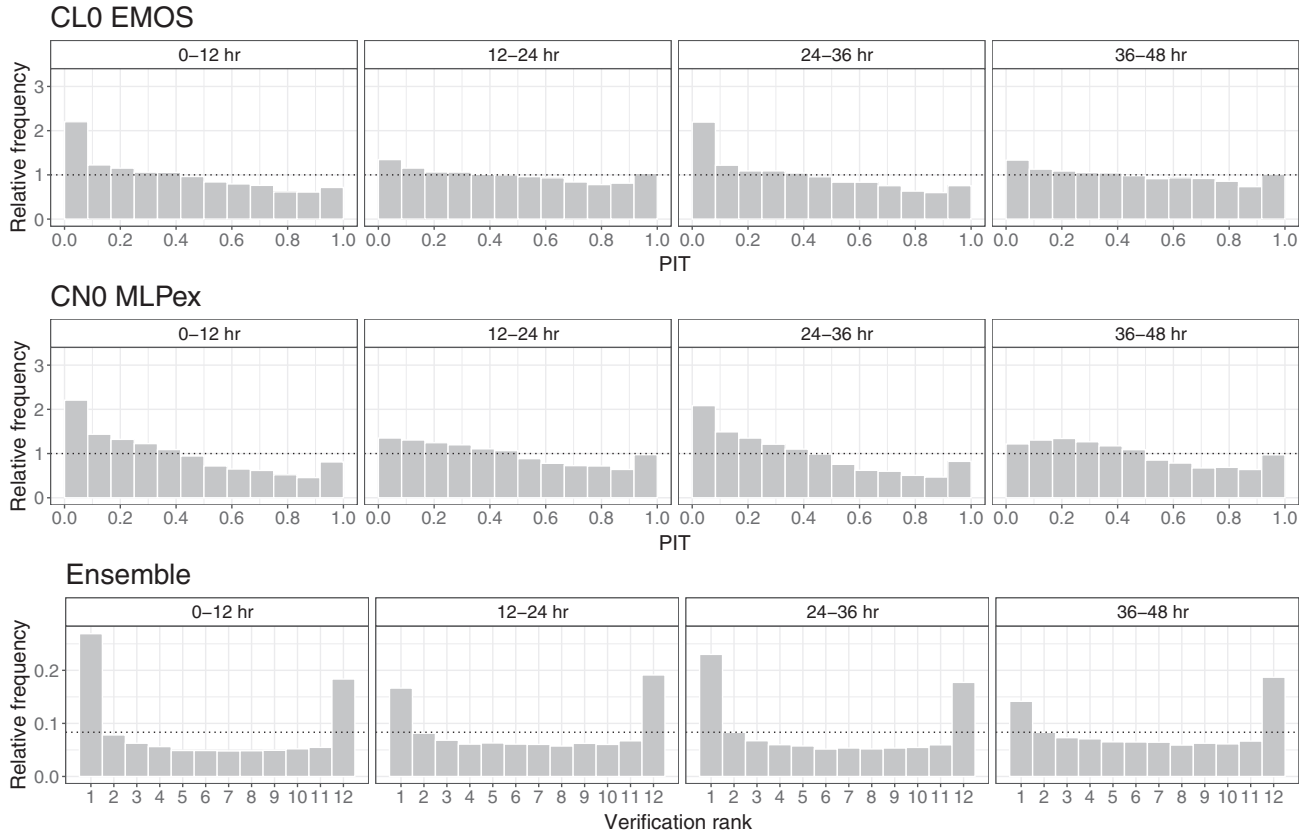

FIGURE 12 Probability integral transform (PIT) histograms of post-processed and verification rank histograms of raw global horizontal irradiance ensemble forecasts for Monor and Nagykőrös for the lead times 0–12, 12–24, 24–36, and 36–48/hr. CL0, censored logistic; CN0, censored normal; EMOS, ensemble model output statistics; MLP, multilayer perceptron; MLPex, extended MLP model.

these additional locations are then validated against observations provided by the solar farm operators, which, as mentioned in Section 2, are in terms of quality far behind the ones provided by the observation network of the HMS. In this way, using the same 1-year verification period from July 1, 2020, to June 30, 2021, one can investigate the robustness of the various approaches. For simplicity, in this section we investigate the performance of the most skilful EMOS (CL0 EMOS) and machine-learning-based (CN0 MLPex) methods only.

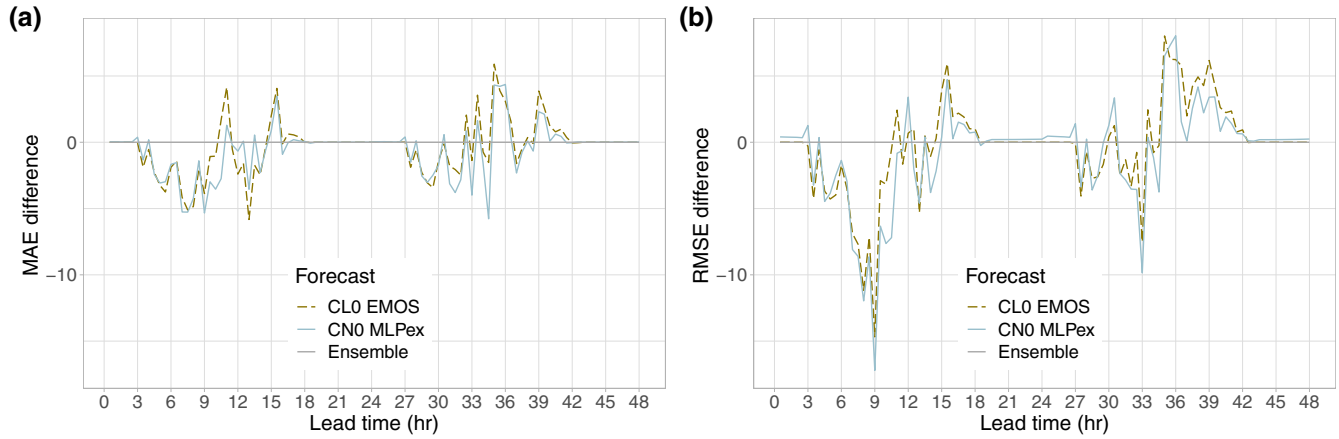

FIGURE 13 Difference in (a) mean absolute error (MAE) of the median forecasts for Monor and Nagykőrös and (b) root-mean-squared error (RMSE) of the mean forecasts from the raw ensemble as functions of the lead time. CL0, censored logistic; CN0, censored normal; EMOS, ensemble model output statistics; MLP, multilayer perceptron; MLPex, extended MLP model. [Colour figure can be viewed at wileyonlinelibrary.com]

Figure 10a again provides the CRPS of the post-processed and raw GHI forecasts, and Figure 10b gives the CRPSS with respect to the raw AROME-EPS. Compared with Figure 6a, the extrapolated models obviously result in smaller improvements in skill. For forecast cases with observed GHI not less than 7.5 $W \cdot m^{-2}$ the mean CRPS values of the CL0 EMOS and CN0 MLPex approaches are 90.39% and 88.91% of the mean CRPS of the AROME-EPS; that is, 7.75pp and 9.80pp higher than the corresponding proportions of Table 3. In general, the machine-learning-based CL0 MLPex approach outperforms the CL0 EMOS model; between 0600 and 1900 UTC it results in a higher skill score in more than 71% of the corresponding lead times.

Figure 11 providing the coverage and average width of the nominal central prediction intervals shows the same general picture as Figure 7. Compared with the raw ensemble, post-processing substantially improves the calibration at the cost of loss in sharpness, although for the calibrated forecasts the deviation from the nominal 83.33% is higher than in Figure 7a. From the two investigated approaches the novel CL0 MLPex results in slightly better coverage and sharper central prediction intervals.

Further, according to the verification rank histograms in the bottom panel of Figure 12, AROME-EPS forecasts for the solar farms at Monor and Nagykőrös are far less underdispersive than for the seven locations used for model training and exhibit a smaller bias (for observed GHI not less than 7.5 $W \cdot m^{-2}$ the average biases of the ensemble median and mean are 14 $W \cdot m^{-2}$ and 17.9 $W \cdot m^{-2}$ respectively). Both post-processing methods investigated result in more uniform, but slightly biased, PIT histograms, especially for lead times 0–12 and 24–36 hr. In this case, for the CL0 MLPex method the α^0_{1234} test rejects uniformity at a 5% level of significance for all available lead times, whereas the CL0 EMOS model has an overall acceptance rate of 32.3%.

Finally, as the differences in MAE of the median forecasts and RMSE of the mean forecasts from the raw ensemble plotted in Figure 13 indicate, for the two solar farms the improvement in the accuracy of point forecasts is not so consistent as before (see Figure 9).

5 | CONCLUSIONS

We propose a general two-step machine-earning-based approach (MLPex) to calibrating ensemble weather predictions resulting in probabilistic forecasts in the form of full predictive distributions. In the first step, with the help of an MLP neural network and a 1D convolutional neural network we provide improved point forecasts of the investigated weather quantity. These forecasts, augmented with simple statistics of the ensemble predictions, then serve as input features for another neural network resulting in the parameters of the predictive distribution. In two case studies, based on 11-member AROME-EPS forecasts of 100 m wind speed and GHI of the HMS, the forecast skill of the suggested MLPex method is compared with the predictive performance of the parametric approach using a single MLP to estimate parameters of the predictive distribution (MLP-S), with the state-of-the-art EMOS models, and with the raw ensemble forecasts. Only short-range predictions up to 48 hr lead time are considered with 15 min and 30 min temporal resolutions for wind speed and GHI respectively. Both weather quantities investigated are of great importance in renewable energy production, and the forecast lead times considered are of

magnitude of the time steps indicated in the scheduling requirements for power plants.

Our case studies confirm that statistical post-processing consistently improves the calibration of probabilistic forecasts, though at the expense of deterioration in sharpness. All the post-processing methods presented result in lower mean CRPS values, better coverage of the nominal 83.33% central prediction intervals, and PIT histograms far closer to the uniform distribution than the corresponding verification rank histograms of the raw ensemble. The novel two-step MLPex approach exhibits the best overall performance in all situations, followed by the corresponding MLP-S method and the EMOS model. For instance, in terms of the mean CRPS, MLP-S outperforms the matching EMOS approach by 1.78–2.87%, and a further 0.39–0.69% gain can be obtained by the use of the MLPex. In the case of solar irradiance, the robustness of the best-performing EMOS and machine-learning-based approach is also investigated by applying regionally trained models to calibration of forecasts for two external locations not used in the training process. This extrapolation does not change the ranking of the forecasts, though the gap in skill between the raw and post-processed predictions is smaller than in the case when the same stations are used both for training and verification.

A general advantage of the machine-learning-based methods investigated is that they are based on the same type of training data as the corresponding EMOS model. MLPex and MLP-S methods do not require long training periods, as in the approach of Ghazvinian et al. (2021), nor do they require additional covariates, such as forecasts of other weather quantities and/or station-specific data, as suggested by Rasp and Lerch (2018) and Ghazvinian et al. (2022). However, the simple and straightforward opportunity of involving additional input features is not excluded either, providing a possible direction of future research.

Another potential avenue of further studies is to consider multivariate post-processing methods providing temporally consistent forecast trajectories. In this context, the MLPex and MLP-S forecasts can serve as initial independent predictions for two-step approaches, where, after univariate calibration, the temporal dependence is restored with the help of an empirical copula calculated using, for example, the actual ensemble forecasts (ensemble copula coupling; Schefzik et al., 2013) or historical observations (Schaake shuffle; Clark et al., 2004). For a detailed comparison of the state-of-the-art multivariate methods with the help of simulated and real ensemble data, we refer the reader to Lerch et al. (2020) and Lakatos et al. (2023) respectively.

AUTHOR CONTRIBUTIONS

Ágnes Baran: Conceptualization, formal analysis, funding acquisition, investigation, software, validation, visualization, writing-original draft, writing-review & editing. **Sándor Baran:** Conceptualization, investigation, methodology, software, validation, writing-review & editing.

ACKNOWLEDGEMENTS

The work leading to this article was done in part during the visit of Sándor Baran to the Heidelberg Institute for Theoretical Studies in July 2022 as guest researcher. Both authors acknowledge the support of the Hungarian National Research, Development and Innovation Office under grant no. K142849. We also thank Gabriella Szépszó and Mihály Szűcs from the HMS for providing the AROME-EPS data. Last, but not least, we are grateful to the two anonymous reviewers, whose constructive comments helped to improve an earlier version of this article.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are confidential; however, for research purposes they are available on request from the HMS.

ENDNOTE

[1]Broadbent, H. (2021). EU solar power hits new record peak this summer. Retrieved from https://ember-climate.org/insights/research/eu-solar-power-hits-new-record-peak-this-summer/.

ORCID

Ágnes Baran https://orcid.org/0000-0002-8764-1673
Sándor Baran https://orcid.org/0000-0003-1035-004X

REFERENCES

Baran, S. & Baran, Á. (2021) Calibration of wind speed ensemble forecasts for power generation. *Időjárás*, 125, 609–624.

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

Baran, S., Szokol, P. & Szabó, M. (2021) Truncated generalized extreme value distribution based EMOS model for calibration of wind speed ensemble forecasts. *Environmetrics*, 32, e2678.

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

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

Buizza, R. (2018b) Ensemble forecasting and the need for calibration. In: Vannitsem, S., Wilks, D.S. & Messner, J.W. (Eds.)

Statistical postprocessing of ensemble forecasts. Amsterdam: Elsevier, pp. 15–48.

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

Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B. & Wilby, R. (2004) The Schaake shuffle: a method for reconstructing space–time variability in forecasted precipitation and temperature fields. *Journal of Hydrometeorology*, 5, 243–262.

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

Demaeyer, J., Bhend, J., Lerch, S., Primo, C., Van Schaeybroeck, B., Atencia, A. et al. (2023) The EUPPBench postprocessing benchmark dataset v1.0. *Earth System Science Data*, 15, 2635–2653.

Fundel, V.J., Fleischhut, N., Herzog, S.M., Göber, M. & Hagedorn, R. (2019) Promoting the use of probabilistic weather forecasts through a dialogue between scientists, developers and end-users. *Quarterly Journal of the Royal Meteorological Society*, 145, 210–231.

Ghazvinian, M., Zhang, Y., Hamill, T.M., Seo, D.-J. & Fernando, N. (2022) Improving probabilistic quantitative precipitation forecasts using short training data through artificial neural networks. *Journal of Hydrometeorology*, 23, 1365–1382.

Ghazvinian, M., Zhang, Y., Seo, D.-J., He, M. & Fernando, N. (2021) A novel hybrid artificial neural network–parametric scheme for postprocessing medium-range precipitation forecasts. *Advances in Water Resources*, 151, 103907.

Gneiting, T. (2011) Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106, 746–762.

Gneiting, T., Balabdaoui, F. & Raftery, A.E. (2007) Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 69, 243–268.

Gneiting, T. & 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. & Goldman, T. (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133, 1098–1118.

Goodfellow, I., Bengio, Y. & Courville, A. (2016) *Deep learning*. Cambridge: MIT Press.

Hemri, S., Scheuerer, M., Pappenberger, F., Bogner, K. & Haiden, T. (2014) Trends in the predictive performance of raw ensemble weather forecasts. *Geophysical Research Letters*, 41, 9197–9205.

IRENA. (2022) *Renewable energy statistics 2022*. Abu Dhabi: The International Renewable Energy Agency.

Jávorné Radnóczi, K., Várkonyi, A. & Szépszó, G. (2020) On the way towards the AROME nowcasting system in Hungary. *ALADIN-HIRLAM Newsletter*, 14, 65–69.

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

Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M. & Inman, D.J. (2021) 1D convolutional neural networks and applications: a survey. *Mechanical Systems and Signal Processing*, 151, 107398.

Knüppel, M. (2015) Evaluating the calibration of multi-step-ahead density forecasts using raw moments. *Journal of Business & Economic Statistics*, 33, 270–281.

Lakatos, M., Lerch, S., Hemri, S. & Baran, S. (2023) Comparison of multivariate post-processing methods using global ECMWF ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 149, 856–877.

Lerch, S., Baran, S., Möller, A., Groß, J., Schefzik, R., Hemri, S. et al. (2020) Simulation-based comparison of multivariate ensemble post-processing methods. *Nonlinear Processes in Geophysics*, 27, 349–371.

Pinson, P. & Messner, J.W. (2018) Application of Postprocessing for renewable energy. In: Vannitsem, S., Wilks, D.S. & Messner, J.W. (Eds.) *Statistical postprocessing of ensemble forecasts*. Amsterdam: Elsevier, pp. 241–266.

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

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

Scheuerer, M. (2014) Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quarterly Journal of the Royal Meteorological Society*, 140, 1086–1096.

Schultz, B., El Ayari, M., Lerch, S. & Baran, S. (2021) Post-processing numerical weather prediction ensembles for probabilistic solar irradiance forecasting. *Solar Energy*, 220, 1016–1031.

Schultz, B. & Lerch, S. (2022) Machine learning methods for postprocessing ensemble forecasts of wind gusts: a systematic comparison. *Monthly Weather Review*, 150, 235–257.

Thorarinsdottir, T.L. & 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.

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

Wilks, D.S. (2018) Univariate ensemble forecasting. In: Vannitsem, S., Wilks, D.S. & Messner, J.W. (Eds.) *Statistical postprocessing of ensemble forecasts*. Amsterdam: Elsevier, pp. 49–89.

Wilks, D.S. (2019) *Statistical methods in the atmospheric sciences*, 4th edition. Amsterdam: Elsevier.

Wind Europe. (2022) Wind energy in Europe 2021. Statistics and the outlook for 2022-2026. https://windeurope.org/intelligence-platform/product/wind-energy-in-europe-2021-statistics-and-the-outlook-for-2022-2026/

Yuen, R.A., Baran, S., Fraley, C., Gneiting, T., Lerch, S., Scheuerer, M. et al. (2018) R package ensembleMOS, version 0.8.2: ensemble model output statistics. https://cran.r-project.org/package=ensembleMOS

How to cite this article: Baran, Á. & Baran, S. (2024) A two-step machine-learning approach to statistical post-processing of weather forecasts for power generation. *Quarterly Journal of the Royal Meteorological Society*, 150(759), 1029–1047. Available from: https://doi.org/10.1002/qj.4635