Solar Energy 307 (2026) 114361

Contents lists available at ScienceDirect

Solar Energy

journal homepage: www.elsevier.com/locate/solener

Post-processing of ensemble photovoltaic power forecasts with distributional and quantile regression methods

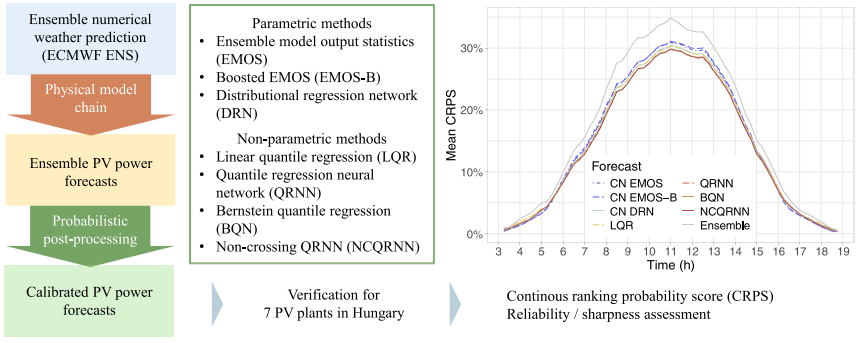
Martin János Mayer[a,b,*], Ágnes Baran[c], Sebastian Lerch[d,e], Nina Horat[f], Dazhi Yang[g], Sándor Baran[c]

[a] Department of Energy Engineering, Faculty of Mechanical Engineering, Budapest University of Technology and Economics, Műegyetem rkp. 3, Budapest, H-1111, Hungary
[b] MTA-BME Lendület "Momentum" Renewable Energy Systems Research Group, Műegyetem rkp. 3, Budapest, H-1111, Hungary
[c] Faculty of Informatics, University of Debrecen, Debrecen, Hungary
[d] Department of Mathematics and Computer Science, Marburg University, Marburg, Germany
[e] Heidelberg Institute for Theoretical Studies, Heidelberg, Germany
[f] Institute of Statistics, Karlsruhe Institute of Technology, Karlsruhe, Germany
[g] School of Electrical Engineering and Automation, Harbin Institute of Technology, Harbin, Heilongjiang, China

HIGHLIGHTS

- Seven statistical post-processing methods are compared for ensemble PV power forecasts.
- Ensemble weather forecasts are converted to PV power using physical model chains.
- All methods significantly improve the reliability of the ensemble forecasts.
- Statistical calibration reduces CRPS by 11.1–14.7% compared to the raw ensemble.
- Quantile regression neural networks outperform distributional regression methods.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Distributional regression network
Ensemble forecast
Ensemble model output statistics
Photovoltaic energy
Post-processing
Quantile regression

ABSTRACT

Accurate and reliable forecasting of photovoltaic (PV) power production is crucial for grid operations, electricity markets, and energy planning, as solar systems now contribute a significant share of the electricity supply in many countries. PV power forecasts are often generated by converting forecasts of relevant weather variables to power forecasts via a model chain. The use of ensemble simulations from numerical weather prediction models results in probabilistic PV forecasts in the form of a forecast ensemble. However, weather forecasts often exhibit systematic errors that propagate through the model chain, leading to biased and/or uncalibrated PV power forecasts. These deficiencies can be mitigated by statistical post-processing. Using PV production data and corresponding short-term PV power ensemble forecasts at seven utility-scale PV plants in Hungary, we systematically evaluate and compare seven state-of-the-art methods for post-processing PV power forecasts. These include both parametric and non-parametric techniques, as well as statistical and machine learning-based approaches. Our results show that compared to the raw PV power ensemble, any form of statistical post-processing significantly improves

* Corresponding author at: Department of Energy Engineering, Faculty of Mechanical Engineering, Budapest University of Technology and Economics, Műegyetem rkp. 3, Budapest, H-1111, Hungary.
Email address: mayer@energia.bme.hu (M.J. Mayer).

https://doi.org/10.1016/j.solener.2026.114361
Received 26 September 2025; Received in revised form 10 January 2026; Accepted 16 January 2026
Available online 29 January 2026

the predictive performance reducing the mean continuous ranked probability score (CRPS) by 11.1–14.7%. Non-parametric methods outperform parametric models, with advanced nonlinear quantile regression models showing the best results. Furthermore, machine learning-based approaches surpass their traditional statistical counterparts by around 2 percentage points in terms of the improvement in mean CRPS over the raw forecasts.

1. Introduction

The global shift toward low-carbon energy systems has brought renewable energy sources to the forefront of electricity production [1,2]. Among these, solar photovoltaic (PV) power has emerged as a key pillar of sustainable energy strategies due to its scalability, declining costs, and widespread deployment, with PV systems now contributing a significant share of electricity supply in many countries [3]. However, the inherently intermittent nature of solar energy presents substantial difficulties, particularly for maintaining the stability and efficiency of power systems. Accurate and reliable forecasting of PV electricity production thus plays a critical role in grid operations, electricity markets, and energy planning. In recent years, a growing emphasis has been placed on probabilistic forecasting approaches [4,5]. These methods move beyond single-valued point forecasts by providing comprehensive predictive information via prediction intervals, quantiles, or full probability distributions, and thus enabling the quantification of forecast uncertainty [6,7].

PV power forecasts are often generated following a three-stage framework, where weather forecasts of global horizontal irradiance (GHI) and other variables from a weather forecasting model are converted to PV power forecasts via a model chain [8–10]. The weather forecasts, which serve as key inputs, are nowadays usually based on numerical weather prediction (NWP) models, which describe physical processes in the atmosphere via systems of differential equations. Ensemble simulations from NWP systems with varying initial conditions or model physics enable the quantification of forecast uncertainties and serve as a straightforward baseline method for generating probabilistic PV power forecasts by applying the PV model chain conversion individually to all ensemble members.

Although this approach allows for propagating forecast uncertainty through the model chain, there is broad evidence that NWP ensemble predictions often show systematic errors [11]. In particular, they are often subject to systematic biases and fail to reliably quantify forecast uncertainty for many variables. Consequently, corrections are achieved through post-processing methods, which rely on statistical or machine learning (ML)-based distributional regression models. These models provide probabilistic forecasts in the form of probability distributions, quantiles, or adjusted ensemble predictions; see [12] for a comprehensive overview of recent developments. A particular focus of recent research has been on the development of modern ML methods. including random forests [13] or neural networks (NNs) [14]. By allowing for the incorporation of multiple meteorological variables as inputs and flexibly modeling nonlinear relationships, they have been found to yield substantial improvements in predictive performance over classical approaches based on statistical methods in various applications, see, e.g., [12,15–17] for overviews and comparisons.

Over the last few years, post-processing methods have also been applied to solar energy prediction, particularly for post-processing NWP forecasts of solar irradiance [4,18–22]. The review by [4] categorized post-processing methods into four groups based on the deterministic or probabilistic nature of the inputs and outputs, namely D2D, D2P, P2D, and P2P. While the paper noted that P2P post-processing was, at that time, the least developed category among the four, and although there has been notable progress in the last four years, P2P post-processing is still less established compared to the deterministic post-processing methods.

In the context of model chain approaches to PV power prediction, P2P post-processing can be applied at different stages, following one of four possible strategies: propagating the raw, unprocessed ensemble weather predictions through the model chain without applying any post-processing; applying post-processing only to the weather inputs before the conversion to PV power; applying post-processing only to the PV power forecasts obtained through the model chain conversion; or applying post-processing both before and after the conversion. [23] compared these strategies using statistical and ML-based post-processing methods based on a benchmark dataset [24] and found that post-processing the PV power predictions is the most promising strategy, which is in line with results from related research on deterministic solar energy prediction [25,26] and on probabilistic wind power forecasting [27]. Furthermore, [23] noted that ML-based post-processing methods outperform their statistical counterparts for solar energy forecasting, albeit by a relatively small margin.

Our overarching aim is to systematically evaluate and compare statistical and ML-based P2P post-processing approaches for the calibration and conditional bias correction of PV power forecasts. Motivated by the findings of [23], we focus on comparing methods for post-processing ensemble forecasts of PV power obtained as the output of the model chain conversion when using raw NWP ensemble predictions as inputs. The main novelty of our work lies in the systematic comparison of a broad set of seven post-processing methods with a particular focus on assessing differences in the predictive performance of parametric distributional regression approaches, which assume a parametric family of probability distributions for the target variable, and non-parametric quantile regression methods, which yield a set of quantiles as their output.

The investigated parametric distributional regression methods include the ensemble model output statistics (EMOS) [28] approach, which is also referred to as non-homogeneous regression and was originally proposed with the assumption of a Gaussian forecast distribution, and has been extended towards solar energy forecasting [20,23]. [29] proposed a gradient boost-based extension of EMOS, which enables the incorporation of additional predictor variables and which we will refer to as EMOS-B or boosted EMOS. A neural network-based distributional regression approach to post-processing was proposed by [14] and will be referred to as the distributional regression network (DRN).

A key drawback of parametric distributional regression approaches is the need to select a suitable parametric family for the conditional distribution of the variable of interest, given the ensemble predictions, which can be a challenge in applications [16]. Non-parametric methods circumvent this disadvantage, with quantile-regression based methods constituting the most popular approach. We here compare standard linear quantile regression to quantile regression neural networks (QRNNs) [30], where neural networks are used to learn non-linear mappings from the input predictors to target quantiles. We further consider Bernstein quantile networks (BQNs) [31], which model the quantile function as a weighted mixture of Bernstein polynomials, as well as the recent non-crossing quantile regression neural network (NCQRNN) approach proposed by [22], which modifies the QRNN architecture to avoid quantile crossing.

Our comparisons are based on a five-year dataset of PV production at seven utility-scale PV plants in Hungary and corresponding ensemble weather forecasts, and thus notably extend the scale of the comparisons conducted in [23] both in terms of the amount of data, as well as the breadth of post-processing methods. Specifically, the novel comparison of parametric and non-parametric approaches allows for assessing the

challenges of choosing a suitable parametric family for PV power production, while considering both classical statistical as well as modern ML-based methods enables insights into the benefits of the potential to flexibly learn non-linear relationships via NNs.

The remainder of this article is organized as follows. Section 2 provides a comprehensive description of the PV power plant and weather forecast data used in this study, as well as the specifics of the model chain that is used for the conversion from weather to PV power forecasts. In Section 3, we introduce the post-processing approaches, provide details of their implementation, and describe the forecast evaluation methods. Section 4 presents the main results, Section 5 summarizes our findings, followed by our conclusions in Section 6. Additional results can be found in the Appendix.

2. Photovoltaic power production and forecast data

The post-processing models are tested for the operational day-ahead power forecasting at seven PV plants in Hungary. The input is a 51-member PV power forecast ensemble, created by converting all members of an ensemble NWP weather forecast into PV power using a physical model chain. The description of the PV plant data, the ensemble NWP, and the model chain are provided in the following subsections.

2.1. Photovoltaic power plant data

The PV power forecasting is performed for seven ground-mounted utility-scale PV plants in Hungary. The locations of the PV plants together with their Köppen–Geier climate classes [32] are shown on the map of Hungary in Fig. 1, and their geographical coordinates and main design parameters are summarized in Table 1. The measured power production data of the PV plants are available for the five full calendar years from 2019 to 2023 with a temporal resolution of 15 min, which fits the operational requirements for scheduling PV plants in Hungary. Only daytime data are considered in this study, selected by a zenith angle $\Theta_z < 90°$ filter, and the daytime data samples with 0 power production are removed from the dataset as they indicate the malfunction or maintenance of the PV plants. The number of valid daytime data samples that are used in the analysis is presented for each year and PV plant in Table 1. Furthermore, to handle the capacity differences of the investigated power plants, both the measured output PV power at a given location and the corresponding PV power forecast are normalized by the nominal AC power of the plant at hand as provided in Table 1.

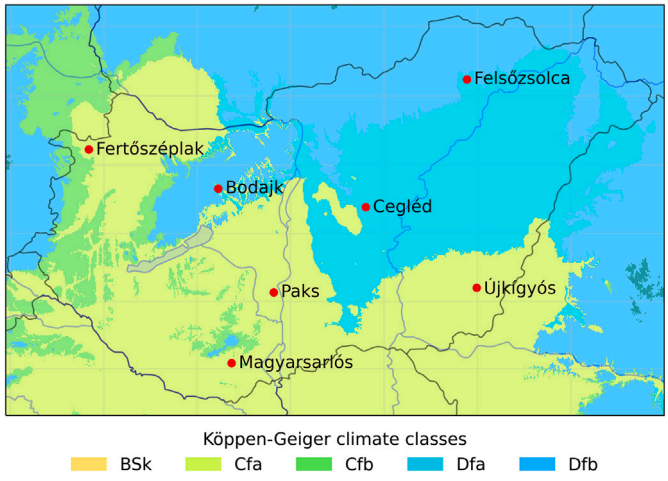


Fig. 1. Locations of the seven utility-scale PV plants considered in this study. The codes refer to the following Köppen–Geier climate classes: BSk – cold semi-arid climate, Cfa – humid subtropical climate, Cfb – subtropical highland climate, Dfa – hot-summer humid continental climate, Dfb – warm-summer humid continental climate.

2.2. Ensemble numerical weather predictions

The weather input data for the PV power forecasts are retrieved from the ensemble (ENS) NWP product of the Integrated Forecasting System (IFS) of the European Centre for Medium-Range Weather Forecasts (ECMWF). The ECMWF ENS is an ensemble of 51 members, including a control forecast and 50 perturbed members. The initial conditions of the NWP, reflecting the current state of the Earth system, are determined by a four-dimensional variational data assimilation (4D-Var) method combining the observations and the latest short-range weather forecasts. The control forecast is created from the best available data using the unperturbed models, whereas the perturbed members are calculated from slightly changed initial conditions with slightly modified model parameterizations. The forecasts of all 51 members for the 24–48-h time horizon are taken from the 00 UTC model run, which is the latest model run that fits the operational requirements of day-ahead forecasting in Hungary. In June 2025, the forecasts of the 00 UTC run are made available at 06:44 UTC for the next day,[1] leaving enough time to prepare the PV power forecasts before the gate closure time of the day-ahead market (DAM) of the Hungarian power exchange (HUPX) at 12:00 CET/CEST.

The forecast weather variables include the GHI (ssrd, surface solar radiation downwards), the ambient temperature at 2 meters (t2m), and the wind speed at 10 meters (norm of the v10 and u10 components). The ECMWF ENS had an 18 km spatial resolution until the Cycle 48r1 model upgrade[2] on 27 June 2023 and 9 km afterwards, and the data from the nearest grid point to the plant location are used for each PV plant. The forecasts are available with a 1-h temporal resolution, which were downscaled to a 15-min resolution to fit the requirements of the Hungarian Transmission System Operator for day-ahead scheduling. The ambient temperature and wind speed are downscaled by linear interpolation, whereas clear-sky interpolation is used for the GHI to better retain the natural daily trend of solar radiation. Thereby, the linear interpolation is performed on the clear sky index, calculated as the ratio of the GHI and its clear-sky counterpart, which is obtained from the McClear service [33].

2.3. Physical photovoltaic model chain

The weather forecasts of all ensemble members are converted to PV power forecasts using a physical model chain of the PV plants. The model chain is a series of physical models, each describing an individual phenomenon [34]. The conversion of weather data to PV power is also called solar power curve modeling; more details on the variety of the existing methods can be found in a recent tutorial review [10]. Model chains can be constructed with different accuracy and complexity depending on the number of steps and the component models selected in each step [35]. In this study, we opt for a detailed model chain in order to account for most of the nonlinearities of the energy conversion to provide the most accurate inputs for the post-processing.

A schematic of the model chain implemented in this study, including the considered modeling steps along with their main inputs and outputs, is shown in Fig. 2. A summary of the models and parameters used in the model chain is provided below, while the detailed description of the models can be found in the original publications of the models. The model chain starts with the calculation of the solar zenith and solar azimuth angles using the solar positioning algorithm of [36]. It is followed by separation modeling, where the GHI is decomposed into its beam and diffuse horizontal components using the temporal-resolution cascade YANG model [37], which emerged as one of the best separation models in a recent worldwide review [38]. The model is used with the parameters proposed for cluster 5 in [39], since all locations at hand

[1] https://confluence.ecmwf.int/display/DAC/Dissemination+schedule
[2] https://www.ecmwf.int/en/about/media-centre/news/2023/model-upgrade-increases-skill-and-unifies-medium-range-resolutions

Table 1
Description of the PV plants considered in this study.

Name	Geographical location		Module orientation		Nominal power (kW)		Number of valid 15-min daytime data samples				
	Lat.	Lon.	Tilt	Azim.	DC	AC	2019	2020	2021	2022	2023
Bodajk	47.33°	18.22°	35°	180°	590	498	17,530	17,547	17,579	17,541	17,533
Cegléd	47.19°	19.80°	35°	180°	590	498	17,067	17,015	17,158	17,090	17,058
Felsőzsolca	48.12°	20.89°	35°	180°	20,038	18,640	17,331	17,337	17,432	17,341	16,934
Fertőszéplak	47.61°	16.84°	35°	180°	590	498	17,533	17,583	17,572	17,506	17,505
Magyarsarlós	46.04°	18.37°	25°	160°	601	502	17,616	17,561	17,568	17,550	17,510
Paks	46.57°	18.82°	35°	180°	20,680	19,160	17,213	17,188	17,273	17,052	17,163
Újkígyós	46.60°	20.99°	35°	180°	590	498	17,095	16,924	17,045	17,027	17,066

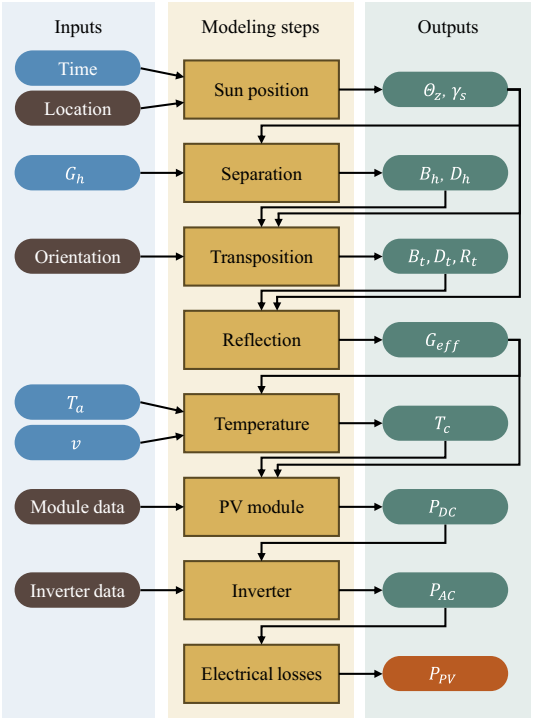


Fig. 2. Schematic of the physical model chain used for converting weather data to PV power.

fall into this one out of the five clusters identified based on cloud cover frequency, aerosol optical depth, and surface albedo climatology.

The next step is to transpose the horizontal irradiance components to the tilted plane of PV arrays. The PEREZ model [40] is selected for this task, which has been widely regarded as the most accurate transposition model for the last more than three decades [41]. The reflection and absorption losses of the PV module cover are accounted for using the angular loss factors proposed by [42]. In addition to the angular loss factor of the beam irradiance that depends on its incidence angle, this model also provides formulae for the sky-diffuse and ground-reflected components. The PV plants at hand feature a mounting structure layout of multiple parallel rows; therefore, the shading losses are estimated by assuming a 2D shading geometry, considering the nonlinear response of the PV power production to the shaded area, as described in [34]. The diffuse irradiance masking caused by the adjacent rows is also taken into account by a reduced sky view factor [43].

The temperature of the solar cells is calculated using the MATTEI model, which includes both the heating of the modules due to the absorbed radiation and the effect of wind speed on the heat transfer coefficient [44]. The power production of the PV modules is modeled using the 5-parameter single-diode equivalent circuits of the modules, as described by [45]. First, the parameter values at nominal conditions are determined based on the datasheet of the PV modules, then they are corrected for the actual irradiance and cell temperature and used to plot the whole I-V characteristic curve of the modules for each timestep. The current, voltage, and power production of the PV modules are obtained from the maximum power point of the characteristic curve. The degradation of the module is accounted for by an initial 2% light-induced loss and a 0.5%/a annual loss factor.

The input power and voltage of the inverter are calculated considering the string layout of the modules and the DC cable losses. The inverter efficiency is estimated using the DRIESSE model [46] as a function of the input power and voltage with parameters fitted to the efficiency values obtained from the datasheet of the inverters. The clipping losses are also accounted for by maximizing the power production at the nominal AC power of the inverter. Finally, the power is summarized for all inverters, and electrical losses on the AC cables, transformers, and other components are deducted to find the power fed into the grid by the PV plant.

3. Post-processing methods and forecast evaluation

In the following sections let $f_1, f_2, \ldots, f_{51} \in [0,1]$ denote the 51-member normalized PV energy ensemble forecast of a given lead time for a given time point and PV power plant, where $f_1 = f_{\mathrm{CTRL}}$ represents the control forecast, while the 50 statistically indistinguishable, therefore exchangeable ensemble members, $f_2, f_3, \ldots, f_{51}$ are also referred to as $f_{\mathrm{ENS},1}, f_{\mathrm{ENS},2}, \ldots, f_{\mathrm{ENS},50}$. Furthermore, denote by $\overline{f}_{\mathrm{ENS}}$ and S^2_{ENS} the mean and variance of these 50 exchangeable ensemble members, respectively, that is,

$$\overline{f}_{\mathrm{ENS}} := \frac{1}{50}\sum_{k=1}^{50} f_{\mathrm{ENS},k} \quad \text{and} \quad S^2_{\mathrm{ENS}} := \frac{1}{49}\sum_{k=1}^{50}\left(f_{\mathrm{ENS},k} - \overline{f}_{\mathrm{ENS}}\right)^2.$$

The general descriptions of the considered parametric and non-parametric models are provided in Sections 3.1 and 3.2, respectively, whereas further implementation details that are common to multiple models are specified in Section 3.4.

3.1. Parametric methods

As mentioned in the Introduction, parametric post-processing methods result in full predictive distributions, and in the EMOS and DRN approaches building on a single parametric law, the chosen distribution family strongly depends on the properties of the predictable quantity. Temperature is mainly considered Gaussian (see, e.g., [14,28]), wind speed follows a skewed distribution with non-negative support, such as truncated normal [47] or log-normal [48], while the positive probability of observing zero precipitation can be handled by left-censoring a skewed distribution, such as generalized extreme value [49] or shifted gamma [50] from below at zero. The same idea led to parametric post-processing models for solar irradiance based on left-censored logistic or Gaussian (see, e.g., [20,21]) laws. However, in the case of PV power, the support of the predictive distribution has a natural upper bound induced by the maximum capacity of the given solar plant. Following [23], the predictive distribution in the EMOS and DRN models detailed in Sections 3.1.1 and 3.1.2, respectively, follows a doubly censored

Gaussian assigning point masses to zero and one, i.e., to both ends of the interval of possible (normalized) PV power values.

3.1.1. Ensemble model output statistics

Consider a Gaussian distribution $\mathcal{N}_0^1(\mu,\sigma^2)$ with location μ and scale σ left-censored at zero and right-censored at one, characterized by the cumulative distribution function (CDF)

$$G(x|\mu,\sigma) := \begin{cases} 0, & x < 0, \\ \Phi\left(\frac{x-\mu}{\sigma}\right), & 0 \le x \le 1, \\ 1, & x > 1, \end{cases} \tag{1}$$

where Φ denotes the CDF of the standard Gaussian law. This distribution assigns masses $p_{\mathrm{LB}} := G(0|\mu,\sigma) = \Phi(-\mu/\sigma)$ to the origin and $p_{\mathrm{UB}} := 1 - G(1|\mu,\sigma) = \Phi((\mu-1)/\sigma)$ to one, while the p-quantile q_p, $(0 < p < 1)$ of (1) equals 0, if $p \le p_{\mathrm{LB}}$, the solution of $G(q_p|\mu,\sigma) = p$, if $p_{\mathrm{LB}} < p < 1 - p_{\mathrm{UB}}$, and 1, if $p \ge 1 - p_{\mathrm{UB}}$.

The parameters of our censored normal (CN) EMOS predictive distribution for (normed) PV power are expressed as the following functions of the (normed) ensemble members

$$\mu = \alpha_0 + \alpha_1 f_{\mathrm{CTRL}} + \alpha_2 \overline{f}_{\mathrm{ENS}} \quad \text{and} \quad \sigma = \exp\left(\beta_0 + \beta_1 \log S^2_{\mathrm{ENS}}\right). \tag{2}$$

Following the optimum score approach suggested by [51], model parameters $\alpha_0, \alpha_1, \alpha_2, \beta_0, \beta_1 \in \mathbb{R}$ are estimated by optimizing the mean value of a proper verification score over training data comprising past forecast-observation pairs. The most popular choices are the ignorance score, which is the negative logarithm of the predictive probability density function (PDF) evaluated at the verifying observation (see, e.g., [Section 9.5.3][52]), which leads to the maximum likelihood estimates, and the continuous ranked probability score (CRPS), defined in Section 3.3. In the case study of Section 4, we utilize the latter.

While EMOS models provide a computationally simple yet powerful tool for statistical post-processing, the rigid functional form of the link functions connecting the ensemble forecasts to the distributional parameters generally does not offer a straightforward way of including additional covariates such as forecasts of related weather quantities or location-specific data like geographical coordinates, altitude, or land use. Moreover, too many predictors can easily lead to overfitting, thereby deteriorating the forecast performance. To circumvent this problem, [29] introduced a boosting algorithm that automatically selects the most important predictors in a nonhomogeneous regression model and provides the maximum likelihood estimates of the corresponding parameters. An implementation of the proposed approach for censored Gaussian predictive distribution can be found in the `R` package `crch` [53]. In the case study of Section 4, as possible predictors, we consider f_{CTRL}, $\overline{f}_{\mathrm{ENS}}$, and S_{ENS}; however, in contrast to the CN EMOS model (2), these covariates might appear both in the location and the scale parameter of the predictive distribution.

3.1.2. Distributional regression network

Distributional regression networks (DRN), introduced by [14], provide an estimation of the parameters of the doubly censored normal predictive distribution by optimizing the mean of the corresponding CRPS over the training data as the loss function of a feedforward neural network. In contrast to EMOS models, considering an extended set of input variables with additional features is straightforward. Further, the large variability of the possible network structures and hyperparameters enables a more flexible post-processing method where non-linear relations are learned in an automated, data-driven manner. A drawback of the DRN models is that, due to the large number of weights, the training typically requires a larger amount of training data. However, in the present case study, the training period is long enough to use the same data as in the case of the EMOS models; for details, see Section 3.4.

3.2. Non-parametric approaches

The non-parametric post-processing methods considered here represent the predictive CDF F by its τ-quantiles, $q_\tau(F) := F^{-1}(\tau) := \inf\{y : F(y) \ge \tau\}$, which are estimated by quantile regression (QR). The most widely used loss function for quantile regression is an asymmetric linear loss called the pinball or quantile loss, which is defined as

$$\rho_\tau(u) := \begin{cases} u\tau, & \text{if } u \ge 0, \\ u(\tau - 1), & \text{if } u < 0, \end{cases} \tag{3}$$

and minimized by $q_\tau(F)$. The quantile loss is not differentiable at $u = 0$, which may cause issues with convergence in the gradient-based optimization methods used for training neural networks. A remedy is to use the quantile Huber loss [54], where the u in (3) is replaced by the Huber norm $h(u)$, calculated as

$$h(u) := \begin{cases} \frac{u^2}{2\epsilon}, & \text{if } |u| \le \epsilon, \\ |u| - \frac{\epsilon}{2}, & \text{if } |u| > \epsilon. \end{cases} \tag{4}$$

If a small value is selected for the ϵ threshold, e.g., $\epsilon = 10^{-8}$, the quantile Huber loss function closely approximates the quantile loss while being differentiable everywhere.

3.2.1. Linear quantile regression

Linear quantile regression (LQR) [55] approximates the quantile q_τ as a linear combination of the predictors. For the calibration of the ensemble forecasts at hand, the predictors are the raw ensemble members, and the calibrated quantile is calculated as

$$q_\tau = \beta_0 + \sum_{k=1}^{51} \beta_k f_k \tag{5}$$

where $\beta_0, \beta_1, \ldots, \beta_{51} \in \mathbb{R}$ are the regression coefficients, fitted to minimize the quantile loss.

While LQR offers a simple solution for approximating the quantiles of the predictive distribution, on the one hand, it requires a separate regression model for each investigated quantile; on the other hand, the predicted quantiles do not necessarily form a nondecreasing sequence (quantile crossing).

3.2.2. Quantile regression neural network

Quantile regression neural networks (QRNN) use a neural network to provide a nonlinear mapping between the predictors and the output quantiles [30]. The QRNN implemented in this study is a feed-forward multilayer perceptron neural network with 51 output neurons, each assigned a quantile loss function with different τ values. In this way, a single QRNN model can estimate all required quantiles simultaneously, eliminating the need for training separate models for each quantile, in contrast to the LQR.

To ensure the optimal fit of the model, an early stopping routine is applied. In this, a validation set is separated from the training data, which is not used directly to adjust the parameters of the model, but the loss function is evaluated for the validation set after each epoch. The training terminates when the validation loss stops improving for a pre-defined number of epochs called the early stopping patience. The convergence and accuracy of the QRNN highly depend on the selection of the hyperparameters, of which the most important are the number of hidden layers and the number of neurons in each hidden layer, the activation function, the learning rate, the early stopping patience, and the batch size.

Note that the problem of quantile crossing can still appear with this approach as well. Moreover, compared to the LQR, QRNN has a much larger number of neuron weights to be estimated, thus requiring far more training data, which is a common drawback of the two other quantile regression methods introduced below.

3.2.3. Bernstein quantile network

Bernstein quantile networks (BQN), proposed by [31], estimate the whole quantile function as a Bernstein polynomial instead of individual quantiles. A Bernstein polynomial of degree n is a linear combination of $n+1$ Bernstein basis polynomials, and the coefficients of the basis polynomials are calculated as the outputs for a neural network. The loss function of the training is the average quantile loss for a set of equidistant quantile levels. The degree of the Bernstein polynomial is a additional hyperparameter of this method, in addition to those listed for the QRNN.

The method is further adjusted by constraining the coefficients to be nondecreasing, which implies that the quantile function is monotonically increasing [16]. Technically, this is implemented by estimating the differences between the coefficients as non-negative values by using a softplus activation function in the output layer of the neural network. A monotonically increasing quantile function ensures that the forecasts for a higher quantile are always equal to or higher than those of a lower quantile, i.e., $q_{\tau_1} \geq q_{\tau_2}$ for $\tau_1 > \tau_2$, in line with the fact that by definition, a CDF must be monotonically increasing.

3.2.4. Non-crossing quantile regression neural network

Non-crossing quantile regression neural networks (NCQRNN), developed by [22], provide an alternative targeted solution to directly enforce the monotonicity of the CDF. QRNN is extended with an additional hidden layer before the output layer that ensures a non-decreasing mapping between the outputs of the previous layer and the nodes of the output layer. The main advantage of this approach is that it has no requirements on the network structure before the non-crossing layer, and thus it can be integrated into any type of neural network. However, NCQRNN has shown a decent performance even with a multilayer perceptron before the non-crossing layer in [22], therefore, this structure is used in the present study. An additional hyperparameter of NCQRNN over the QRNN is the number of neurons in the non-crossing layer, which must be equal to or higher than the number of output neurons.

3.3. Forecast evaluation

The performance of both probabilistic predictions (a forecast ensemble, a full predictive distribution, or predictive quantiles) and point forecasts (median or mean of the forecast ensemble or predictive distribution) can be evaluated with the help of scoring rules, which are loss functions assigning numerical values to forecast-observation pairs. In the case of the predictive median, we consider the mean absolute error (MAE), while the mean forecasts are evaluated with the help of the root mean squared error (RMSE) and the mean bias error (MBE), also known as the mean error (see, e.g., [Section 9.3.1][52]). Since the MAE is minimized by the median and the mean squared error is minimized by the mean [56], the aforementioned pairing of the point forecast with the error metrics ensures the consistency of the verification, i.e., the deterministic forecasts are only evaluated with metrics that they are optimal for [57,58].

In the case of probabilistic forecasts, one of the most popular scoring rules is the continuous ranked probability score (CRPS) [Section 9.5.1][52], as it is strictly proper and simultaneously addresses the calibration and the sharpness of the forecasts [51]. Calibration indicates a statistical consistency between the probabilistic forecast and the corresponding observation, while sharpness addresses the concentration of the forecasts. When a probabilistic forecast corresponding to an observation $x \in \mathbb{R}$ materializes in the form of a predictive CDF F, the CRPS is defined as

$$\mathrm{CRPS}(F,x) := \int_{-\infty}^{\infty} \Big[F(y) - \mathbb{I}_{\{y \geq x\}}\Big]^2 \mathrm{d}y = \mathsf{E}|X - x| - \frac{1}{2}\mathsf{E}|X - X'|, \tag{6}$$

where $\mathbb{I}_A$ denotes the indicator function of a set A, while X and X' are independent random variables distributed according to F and having a finite first moment. Note that for the doubly censored Gaussian distribution, the CRPS has a closed form [59] that allows efficient estimation of the parameters of the EMOS model presented in Section 3.1.1.

In the case of a forecast ensemble $f_1, f_2, \ldots, f_K$, in (6) the predictive CDF F should be replaced by the empirical CDF $\widehat{F}_K$, resulting in the expression

$$\mathrm{CRPS}(\widehat{F}_K, x) = \frac{1}{K}\sum_{k=1}^{K}\Big|f_k - x\Big| - \frac{1}{2K^2}\sum_{k=1}^{K}\sum_{\ell=1}^{K}\Big|f_k - f_\ell\Big|, \tag{7}$$

see, e.g., [60]. This version of the empirical CRPS is implemented in the `scoringRules` package of `R` [59] and slightly differs from the ensemble CRPS defined in [Section 9.7.3] [52]. The same formula (7) for the CRPS also applies when the predictive distribution is represented by its quantiles.

Furthermore, similar to other strictly proper scoring rules, the CRPS has an algebraic decomposition into a reliability (REL), resolution (RES) and uncertainty (UNC) term

$$\mathrm{CRPS} = \mathrm{REL} - \mathrm{RES} + \mathrm{UNC},$$

where reliability summarizes the calibration of the probabilistic forecast, resolution is closely related to its sharpness, while uncertainty represents the climatological variability and thus depends only on observations [61,62].

In Section 4, the predictive performance of a forecast F for a given time of the day is quantified, among others, with the help of the mean CRPS and the MAE over all forecast cases used for verification. For ranking the different forecasts, we also consider the continuous ranked probability skill score (CRPSS; see, e.g., [51]) and the mean absolute error skill score (MAES), which provide the improvement in mean CRPS and MAE of a forecast F over a reference forecast F_{ref}, and are defined as

$$\mathrm{CRPSS} := 1 - \frac{\overline{\mathrm{CRPS}}_F}{\overline{\mathrm{CRPS}}_{F_{\mathrm{ref}}}} \qquad \text{and} \qquad \mathrm{MAES} := 1 - \frac{\mathrm{MAE}_F}{\mathrm{MAE}_{F_{\mathrm{ref}}}},$$

where $\overline{\mathrm{CRPS}}_F$, MAE_F and $\overline{\mathrm{CRPS}}_{F_{\mathrm{ref}}}$, $\mathrm{MAE}_{F_{\mathrm{ref}}}$ denote mean score values corresponding to forecasts F and F_{ref}, respectively.

Furthermore, to gain insight into the forecast skill of the quantile forecasts, we make use of the quantile score (QS) [Section 9.6.1] [52], defined via the pinball loss (3) as

$$\mathrm{QS}_\tau(F,x) := \rho_\tau\big(x - q_\tau(F)\big). \tag{8}$$

Note that the QS is proper and its integral over all quantiles results in half of the CRPS [63].

Calibration and sharpness of predictive distributions can also be assessed with the help of the coverage and average width of $(1-\alpha)100\%$, $\alpha \in (0,1)$, central prediction intervals (intervals between the lower and upper $\alpha/2$ quantiles of the predictive CDF), respectively. In this context, prediction interval coverage probability (PICP) is the proportion of verifying observations located in the corresponding central prediction interval, which for a calibrated forecast should be around $(1-\alpha)100\%$, while the prediction interval average width (PIAW) quantifies the concentration of the predictive law. Note that when a K-member ensemble forecast is also involved in the study, to ensure fair comparability in the detailed analysis, level α is chosen to match its nominal coverage of $(K-1)/(K+1)100\%$ meaning 96.15% for the 51-member PV forecasts at hand. Moreover, since the prediction interval of interest depends on the application of the forecasts, PIAW is also presented for all possible central prediction intervals defined by the 51-member ensemble as a function of the nominal and empirical coverage rates.

However, PICP alone is not sufficient to evaluate the reliability of probabilistic forecasts, since it may suggest perfect reliability even if both quantiles defining the prediction interval are biased in the same direction [6]. A better approach is to evaluate the reliability at all quantile

levels individually using a reliability diagram that plots the proportion of the observations that are actually smaller than the forecasts for each quantile at hand as a function of the quantile level. The reliability curve of a perfectly calibrated forecast lies close to the diagonal.

Another simple graphical tool for visual assessment of the calibration of probabilistic forecasts given either as a forecast ensemble or as a sample drawn from a predictive distribution is the verification rank histogram [Section 9.7.1] [52]. The verification rank is defined as the rank of the observation with respect to the corresponding forecast, which for a calibrated K-member ensemble should be uniformly distributed on the set $\{1, 2, \ldots, K+1\}$. Bias results in triangular shapes, while ∪- and ∩-shaped rank histograms suggest under- and overdispersion. Moreover, one can also quantify the deviation from the uniform distribution with the help of the reliability index (RI) [64], defined as

$$\mathrm{RI} := \sum_{r=1}^{K+1} \left| \rho_r - \frac{1}{K+1} \right|,$$

where ρ_r is the relative frequency of rank r over all forecast cases in the verification period.

3.4. Implementation details

All 51 members of the ECMWF ensemble forecasts contain GHI, ambient temperature, and wind speed data, which are converted to PV power using a physical model chain. All perturbed NWP ensemble members are generated from randomly issued initial conditions; therefore, there is no continuity between the same-numbered members of different model runs. To that end, the 50 perturbed PV power forecast members are sorted at ascending order in each timestep, which can be seen as converting them to equidistant quantile forecasts, and the sorted ensemble is used as the input for the post-processing models.

In the case of EMOS modeling, all lead times are considered separately, leading to at most 65 distinct models per PV power plant. We consider a fixed training period of 1460 calendar days between 1 January 2019 and 30 December 2022. Note that a rolling training window of the same length has also been tested without providing significantly different results, whereas shorter training periods (365-, 730-, 1096-day have been tested) decrease the forecast skill. As mentioned in Section 3.1.1, the parameters of the doubly censored normal EMOS (CN EMOS) model are estimated by optimizing the mean CRPS over the training data, while in the boosted version of EMOS, referred to as CN EMOS-B, the control member f_{CTRL} and the mean $\overline{f}_{\mathrm{ENS}}$ and standard deviation S_{ENS} of the exchangeable ensemble members are used as covariates.

For CN EMOS-B and all the other post-processing methods, we use the same training period as for the EMOS model, but all the lead times are pooled, resulting in a single trained model for each method and PV plant.

For the doubly censored DRN (CN DRN) model, the neural network is a multilayer perceptron with three hidden layers consisting of 15, 10, and 10 neurons, respectively, all of which use a ReLU activation function. In the output layer, there are two neurons, corresponding to the number of estimated parameters. To ensure the non-negativity of the scale parameter, one of the neurons applies a softplus activation, while the other activation function is linear. The input features are simply the 51-member ensemble for the PV power forecast. To optimize the loss, we apply the Adam optimizer with a learning rate value of 0.001 and use a batch size of 256. An early stopping criterion terminates the training if the loss function value computed on a validation set does not decrease over six consecutive epochs. Following common practice in DRN-based post-processing, we train an ensemble of ten neural networks and average the output parameters [65].

The QRNN, BQN, and NCQRNN models include a similar multilayer perceptron with up to two hidden layers with 5 to 200 neurons each. The considered activation functions are the ReLU, softplus, logistic, and tanh functions, the learning rate is selected from the 0.0005 to 0.05 interval, the early stopping patience may range from 5 to 50 epochs, and the batch size is between 200 and 20,000. The degree of the Bernstein polynomial in BQG ranges from 6 to 15, whereas the number of non-crossing neurons in NCQRNN is between 51 and 60. The optimal hyperparameters for each model and PV plant are selected from the aforementioned intervals/options using the Optuna framework [66]. For this, the training data of four years is divided into five-day-long blocks, and the first three days of each block are used for the actual training of the models, the fourth days are used as validation data for the early stopping, and the fifth days are used to form a holdout dataset. The Optuna hyperparameter optimization studies were run with 100 trials for each model, and the hyperparameter sets resulting in the lowest CRPS for the holdout data are selected. After finding the optimal hyperparameters, reported in Table 3, the data for every fifth day are used for validation, and the rest for the training of the final model.

4. Results

In the following, we present a detailed comparison of the predictive performance of parametric and non-parametric post-processing methods introduced in Sections 3.1 and 3.2, respectively. All models are trained locally; i.e., post-processing models for a given PV power plant are based solely on past forecast-observation pairs for that specific location. For verification, we use power data for the calendar year 2023 and, as mentioned in Section 2.1, consider only daytime forecasts corresponding to positive observed PV power, meaning at most 65 observations/day between 03:00 and 19:00 UTC. Both the parametric and non-parametric post-processing methods are based on normalized data, i.e., both PV power forecasts and PV power production of a plant are normalized by the corresponding nominal AC power provided in Table 1. Furthermore, to ensure a fair comparability of the parametric methods resulting in full predictive distributions, non-parametric techniques providing quantile forecasts, and the raw ensemble, we consider 51 equidistant quantiles from the predictive distributions for each post-processing model. These quantiles are then transformed back to the original scale.

Normalizing by the nominal AC power, as is done with the input and target variables of the model, effectively scales the power values strictly to a 0–1 range; however, it is not the best basis for normalizing error metrics. The nominal AC power of PV plants depends on the inverter sizing factor, which may vary over a wide range without substantially affecting the performance of the PV plants. Therefore, even very similar PV systems (in terms of nominal DC power or annual energy production) can have significantly different nominal AC power and thus AC-normalized error metrics, which can falsely suggest different forecasting performance. Instead, we prefer normalizing the errors to the mean power production, which is directly proportional to the total amount of electricity generated. This offers improved interoperability, e.g., the mean-normalized MAE reflects the ratio of the required balancing energy to the total produced energy. Therefore, the differences between the nominal powers of the seven considered plants are compensated for by reporting score values normalized by the mean daytime power production of the PV plants listed in Table 2.

For an overall evaluation, consider first the mean scores averaged over all PV plants in Table 4 for all post-processed and raw PV forecasts. The lowest CRPS is achieved by the nonlinear QR methods, namely the QRNN, BQN, and NCQRNN, achieving a CRPS reduction of 14.67–14.73% over the raw ensemble (represented by the CRPSS); however, even the least effective EMOS models reach a CRPSS of 11.08%. The reliability-resolution (REL-RES) decomposition reveals that all methods improve the reliability of the forecast at the cost of a decreased resolution (sharpness). Even though all methods are able to improve the reliability substantially, the non-parametric models achieve better reliability values for 0.44–0.83% compared to 1.53–1.97% of the parametric models. However, the less effective calibration of the parametric models is partly compensated for by a slightly higher resolution.

Table 2
Mean daytime power production of the PV plants.

Name	Bodajk	Cegléd	Felsőzsolca	Fertőszéplak	Magyarsarlós	Paks	Újkígyós
Power (kW)	169.27	178.29	5503.95	168.37	166.27	6161.85	179.06

Table 3
Optimal hyperparameters for the QRNN, BQN, and NCQRNN methods.

Model	Hyperparameter	Bodajk	Cegléd	Felsőzsolca	Fertőszéplak	Magyarsarlós	Paks	Újkígyós
QRNN	Activation function	sigmoid	softplus	ReLU	sigmoid	sigmoid	sigmoid	sigmoid
	Learning rate	0.00050	0.01285	0.00386	0.00120	0.00054	0.00322	0.00740
	Early stopping patience	48	47	35	47	30	37	44
	Batch size	284	667	590	222	602	366	306
	Neurons per hidden layer	152/26	158/135	108/125	124/189	29/7	178/90	38/11
BQN	Berstein polynomial degree	10	7	10	15	15	11	10
	Activation function	sigmoid	softplus	tanh	sigmoid	sigmoid	sigmoid	sigmoid
	Learning rate	0.00389	0.00328	0.00134	0.00224	0.00055	0.00317	0.00514
	Early stopping patience	43	47	33	42	47	31	33
	Batch size	409	310	502	496	836	391	377
	Neurons per hidden layer	168/50	142/155	107/85	169/200	176/40	68/94	56/29
NCQRNN	Neurons in non-crossing layer	59	52	51	52	54	55	52
	Activation function	sigmoid	ReLU	softplus	softplus	sigmoid	softplus	sigmoid
	Learning rate	0.00326	0.00319	0.00211	0.00121	0.01088	0.00620	0.00085
	Early stopping patience	48	39	50	42	41	37	47
	Batch size	6425	1318	750	242	2669	200	267
	Neurons per hidden layer	13	80/115	196/16	152/160	118/11	32/11	152/141

Table 4
Summary of CRPS, reliability, and resolution of the probabilistic forecasts as well as the MAE, MBE, and RMSE of the consistently summarized deterministic forecasts averaged for all PV plants.

Forecast	Probabilistic forecast				Median	Mean	
	CRPS	Reliability	Resolution	CRPSS	MAE	MBE	RMSE
CN EMOS	18.95%	1.66%	32.54%	11.13%	26.46%	5.19%	42.41%
CN EMOS–B	18.96%	1.53%	32.40%	11.08%	26.22%	4.16%	42.20%
CN DRN	18.58%	1.97%	**33.22%**	12.85%	26.00%	3.48%	42.14%
LQR	18.65%	0.46%	31.64%	12.56%	26.39%	2.12%	42.03%
QRNN	**18.18%**	**0.44%**	32.09%	**14.73%**	25.84%	1.68%	**41.86%**
BQN	18.19%	0.83%	32.48%	14.69%	**25.78%**	2.07%	41.91%
NCQRNN	18.20%	0.47%	32.10%	14.67%	**25.78%**	**1.65%**	41.91%
Ensemble	21.33%	7.63%	36.14%	0.00%	27.58%	9.79%	44.07%

Table 5
Overall mean CRPS of post-processed and raw PV power forecasts normalized to the mean daytime power production of the PV plants.

Forecast	Bodajk	Cegléd	Felsőzsolca	Fertőszéplak	Magyarsarlós	Paks	Újkígyós
CN EMOS	19.44%	19.01%	20.62%	19.83%	17.57%	17.77%	18.42%
CN EMOS–B	19.40%	18.91%	20.67%	19.97%	17.49%	17.75%	18.56%
CN DRN	18.83%	18.57%	20.20%	19.46%	17.31%	17.46%	18.27%
LQR	19.06%	18.53%	20.31%	19.55%	17.41%	17.45%	18.23%
QRNN	**18.53%**	18.13%	19.67%	**19.02%**	17.10%	17.07%	**17.77%**
BQN	**18.53%**	**18.11%**	**19.63%**	19.05%	17.14%	**17.05%**	17.84%
NCQRNN	18.55%	18.11%	19.78%	19.08%	**17.01%**	17.07%	17.78%
Ensemble	21.99%	21.23%	23.14%	21.87%	21.20%	19.69%	20.17%

To address the dependence of the results on the PV plant locations, the CRPS values are presented individually for each PV plant in Table 5. The conclusions drawn from the mean score values also hold for all PV plants, as there are no significant differences in the order of the methods. The lowest CRPS is consistently achieved by one of the nonlinear QR models, which ended up head-to-head in all locations, with each being the best performer in at least one PV plant. The achieved CRPSS of the best model, however, strongly depends on the location, with the lowest and highest CRPSS values being 11.89% and 19.32% in Újkígyós and Magyarsarlós, respectively.

The mean CRPS of post-processed and raw PV forecasts for each hour of the day is displayed in Fig. 3(a). As confirmed by the skill scores in Fig. 3(b), compared to the raw ensemble, post-processing results in a substantial relative improvement of around 10% during the hours of peak PV power production (06:00–16:00 UTC). In this period of the day, the differences between the competing calibrated forecasts are rather small with the advanced quantile-based methods (QRNN, BQN, NCQRNN) exhibiting the best, almost identical skill, followed by the LQR and CN DRN, whereas the two EMOS variants are slightly behind. This trend is in line with the conclusions drawn from the overall scores

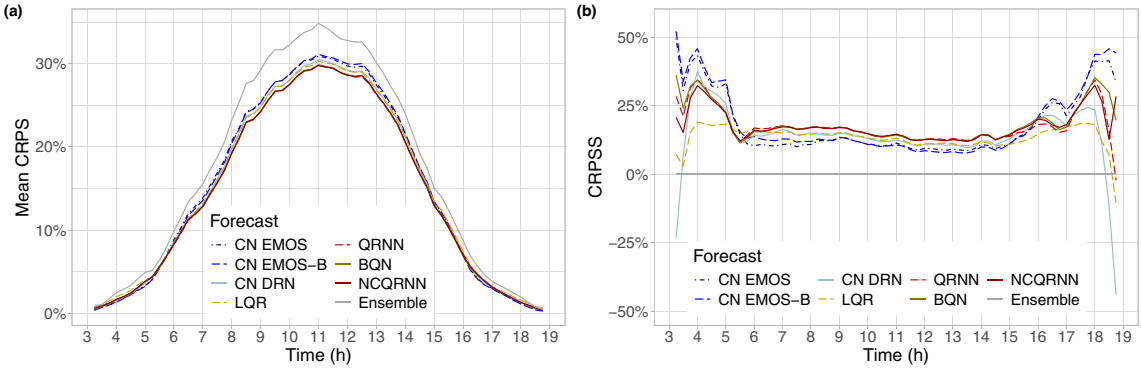


Fig. 3. Mean CRPS of post-processed and raw PV power forecasts normalized to the mean daytime power production of the PV plants (a) and CRPSS of post-processed forecasts with respect to the raw ensemble (b) as functions of the observation time.

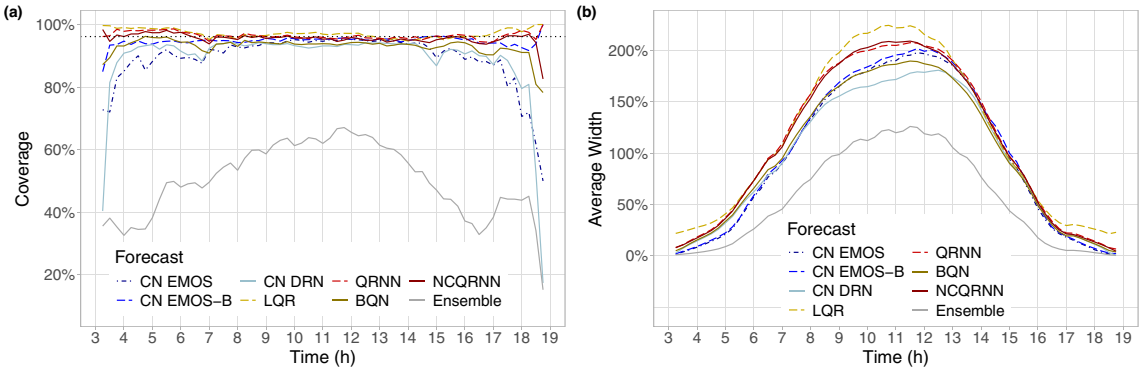


Fig. 4. Coverage (PICP) (a) and average width (PIAW) (b) of nominal 96.15% central prediction intervals of post-processed and raw PV power forecasts normalized to the mean daytime power production of the PV plants as functions of the observation time. In panel (a), the ideal coverage is indicated by the horizontal dotted line.

of Tables 4 and 5. Results on the significance of the differences in mean CRPS among the various forecasts are presented in the Appendix. They confirm that all post-processing methods significantly outperform the raw PV power ensemble, and the advantage of the three quantile-based methods that perform the best over the other four approaches is also significant at a 5% level.

The improved calibration of post-processed forecasts is also clearly visible in the coverage (PICP) values presented in Fig. 4(a). While the maximal PICP of the raw forecasts is just slightly above 67%, between 08:00–14:00 UTC, all post-processing approaches result in almost perfect coverage, which is maintained by the best-performing non-parametric approaches for all observation times, except the most extreme ones. The reliability diagram in Fig. 5 and the rank histograms in Fig. 6 allow a more detailed assessment of reliability over the whole range of probability levels. These diagrams not only confirm the significant underdispersion of the raw ensemble but also reveal that both versions of the EMOS model overcompensate for this, resulting in a slight overdispersion in the medium probability range.

As indicated in Fig. 4(b), the price of the better calibration is the loss in sharpness. Among the competing calibrated forecasts, the CN DRN results in the lowest overall PIAW, followed by the BQN and the two EMOS methods. Fig. 7(a) shows the mean PIAW for all central prediction intervals with different nominal coverage rates, clearly revealing the widening of all prediction intervals as a result of the calibration. That said, this diagram does not account for the fact that the prediction intervals of an uncalibrated ensemble cover a significantly lower proportion of the observations as compared to what their nominal coverage rate suggests, and thus misleadingly imply the deterioration of the

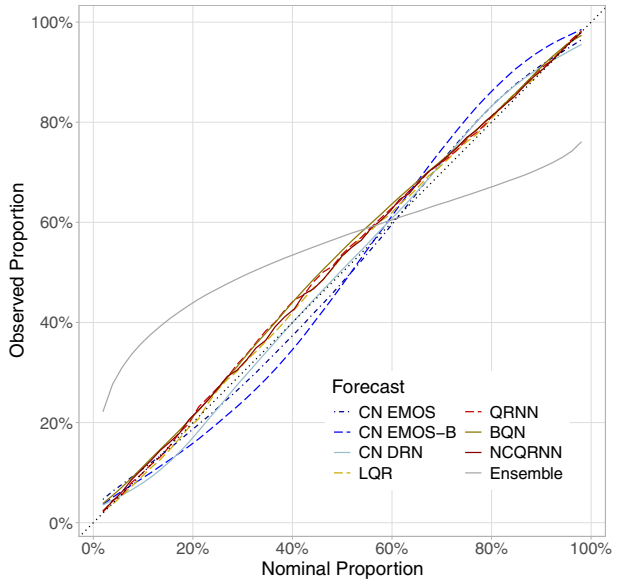


Fig. 5. Reliability diagrams of post-processed and raw PV power forecasts.

forecast quality. To better represent the sharpness of the forecasts with respect to their calibration, Fig. 7(b) plots the mean PIAW as a function of the empirical coverage rate (i.e., the PICP) instead of the nominal one. This novel graphical representation reveals that, compared to the

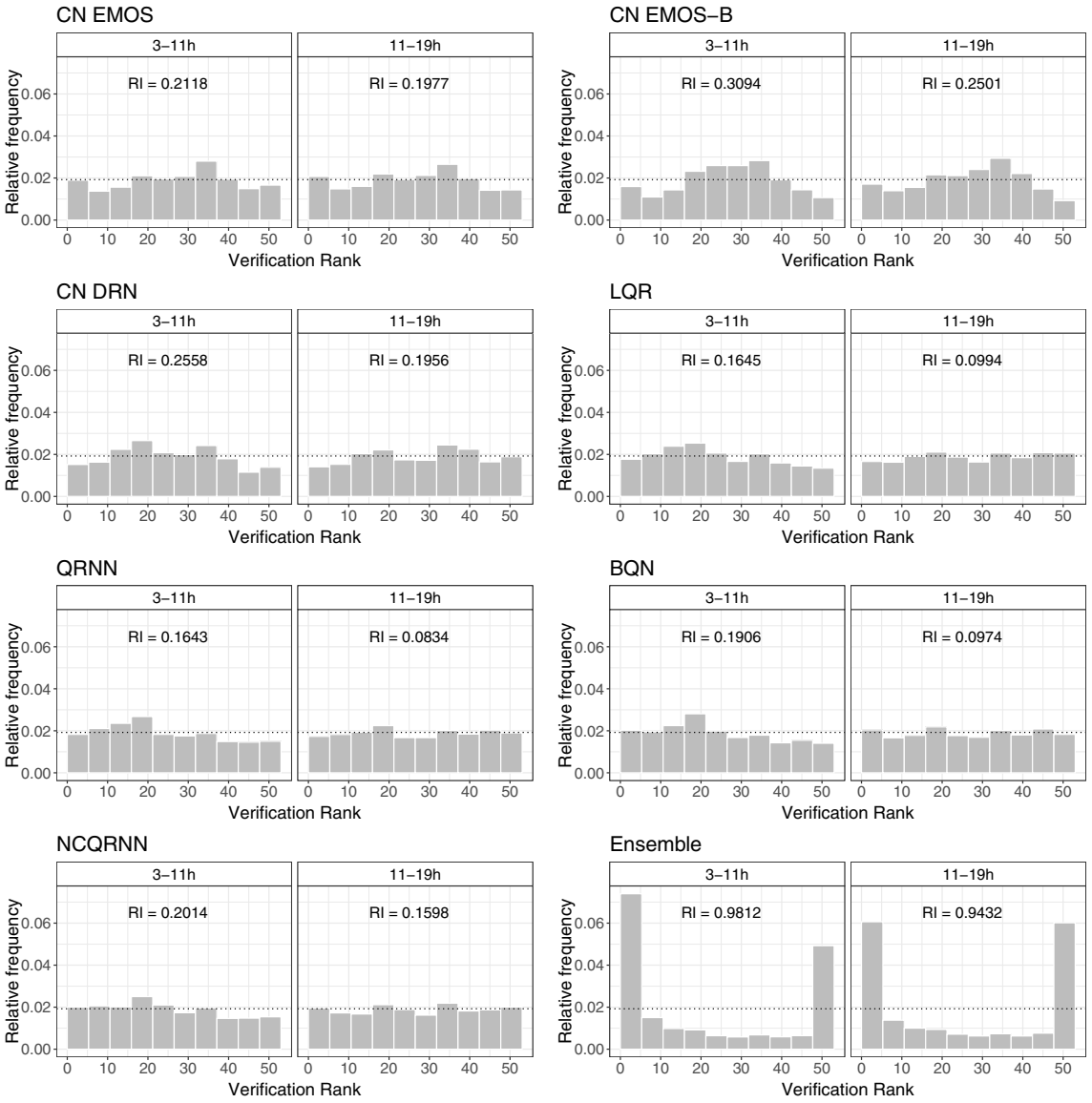


Fig. 6. Verification rank histograms of post-processed and raw PV power forecasts together with the corresponding reliability indices for observation times 3–11 h and 11–19 h.

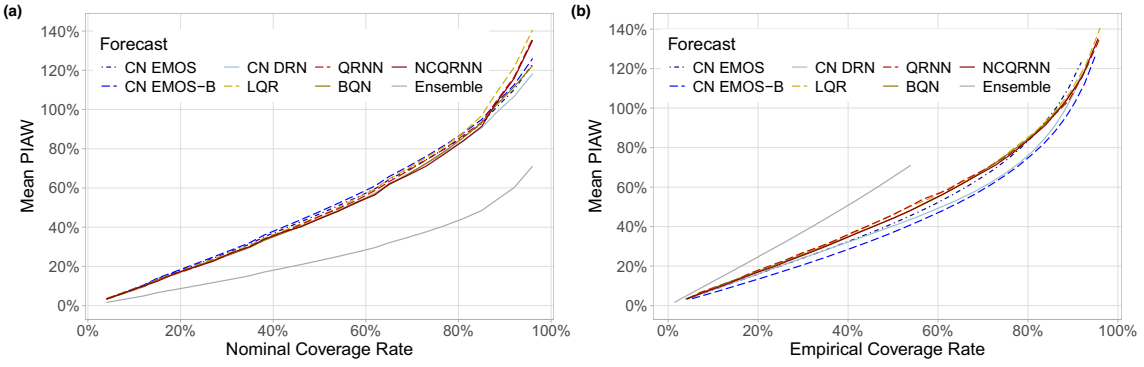


Fig. 7. Sharpness of post-processed and raw PV power forecasts normalized to the mean daytime power production of the PV plants as functions of the nominal coverage (a) and the corresponding empirical coverage (PICP) (b).

proportion of the observations they cover, the calibrated prediction intervals are narrower compared to the raw ensemble, in line with the improved CRPS.

Considering that the CRPS is double the integral of the QS over all quantiles [63,67], the QS diagram in Fig. 8 can reveal which quantile levels contribute the most to the CRPS improvement. Since, according to

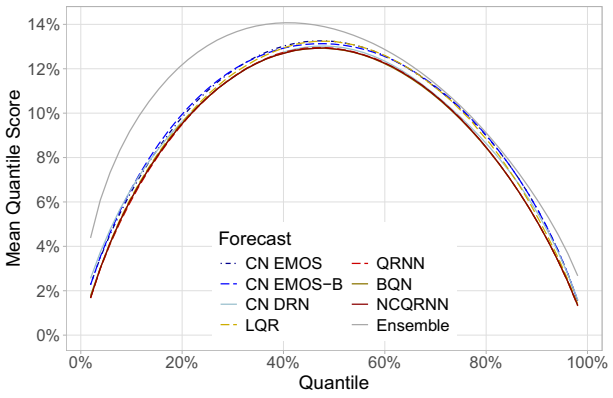


Fig. 8. Quantile score of post-processed and raw PV power forecasts normalized to the mean daytime power production of the PV plants.

the reliability diagram in Fig. 5, the raw ensemble is only reliable around the 62% probability level, the biggest improvement in QS is achieved at the lower quantile range. Comparing the different post-processing methods, the QRNN, BQN, and NCQRNN have almost the same QS for all quantiles, which is slightly but consistently lower than the QS of other methods. The LQR performs well up to the 30% and above the 90% quantile, but lags behind in between. In contrast, the DRN catches up with the nonlinear QR methods between 20% and 70%, but slightly underperforms around the extreme probability levels.

In terms of the accuracy of point forecasts, the average error metrics for all PV plants are presented in Table 4. The greatest relative MAE improvement over the raw ensemble is 6.51%, achieved by either the BQN or NCQRNN models, whereas the greatest RMSE improvement is 5.02%, achieved by the QRNN. The raw forecasts have a significant positive MBE, originating from both the ensemble NWP and the model chain. All methods decrease the MBE, but the non-parametric approaches can provide a better correction as compared to the parametric models, among which the simple EMOS retains more than half of the original bias.

According to Fig. 9, showing the MAE of the median of probabilistic forecasts and the MAES with respect to the raw ensemble, while the ranking of the various forecasts is slightly different, as CN DRN can often catch up with the best-performing non-parametric methods, the (normalized) improvement in peak hours (06:00–16:00 UTC) hardly exceeds 12%. For results addressing the significance of differences in MAE among the various forecasts, we again refer to the Appendix. A similar behaviour can be observed for the daily evolution of the RMSE of the mean forecasts (not shown). Overall, the gain of statistical post-processing in deterministic forecasting is not so striking, but still remarkable considering that the goal of the post-processing is probabilistic calibration and only the raw ensemble members are used as predictors.

Finally, time series plots of the raw and all post-processed forecasts are shown for a sample week in Fig. 10. There is no significant visual difference between the models within both the parametric and non-parametric model categories, but forecasts created by these two different approaches can be clearly distinguished. The main difference is that the parametric models assign narrow prediction intervals to the periods with supposedly clear skies (see the mornings of the 20th, 24th, and 27th of April), whereas the non-parametric methods assign much wider prediction intervals towards the lower PV power values, especially for the high coverage rates. In this way, these models give some probability to the events when the clear sky forecasts are wrong, which explains both the improved reliability and lower sharpness.

Overall, the results show that the nonlinear QR methods, namely the QRNN, BQN, and NCQRNN, are consistently the best performers in almost all respects. The lower performance of the LQR can be justified by its linearity, while for the parametric method, the pre-defined shape of the CDF limits the performance. Among the three nonlinear QR methods, the simple QRNN has a slight edge over the others, suggesting that the least constrained method can yield the best results in this application. However, one should note that a large historical dataset covering four full calendar years was available to train the models. When less data are available for training, constraints, for instance, on the CDF, can prove effective to avoid nonphysical results, and this can be the application where parametric methods excel.

5. Discussion

The present work provides a detailed comparison of seven state-of-the-art approaches to statistical post-processing of 51-member PV power ensemble forecasts obtained from operational ECMWF ensemble weather forecasts as outputs of site-specific model chains. With the help of PV power data from seven PV plants in Hungary, we evaluated the skill of the doubly censored Gaussian ensemble model output statistics (CN EMOS) model, its boosted version (CN EMOS-B) and the corresponding distributional regression network (CN DRN) technique together with the linear quantile regression (LQR), quantile regression neural network (QRNN) and its non-crossing variant (NCQRNN) and Bernstein quantile network (BQN) methods.

We found that compared to the raw PV power ensemble, any form of statistical post-processing significantly improves the predictive performance, resulting in, for instance, an 11.08–14.73% overall gain in terms of the mean continuous ranked probability score. Post-processing also decreases the quantile score, the mean absolute error of the median and the root mean squared error of the mean, improves the reliability, and yields almost perfect coverage of the nominal central prediction intervals; however, at the cost of a deterioration in sharpness.

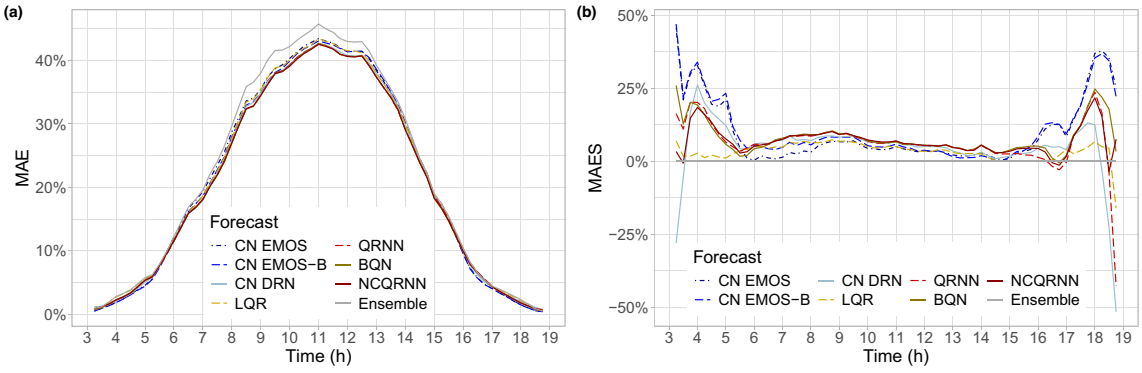


Fig. 9. MAE of the median of post-processed and raw PV power forecasts normalized to the mean daytime power production of the PV plants (a) and MAES of post-processed forecasts with respect to the raw ensemble (b) as functions of the observation time.

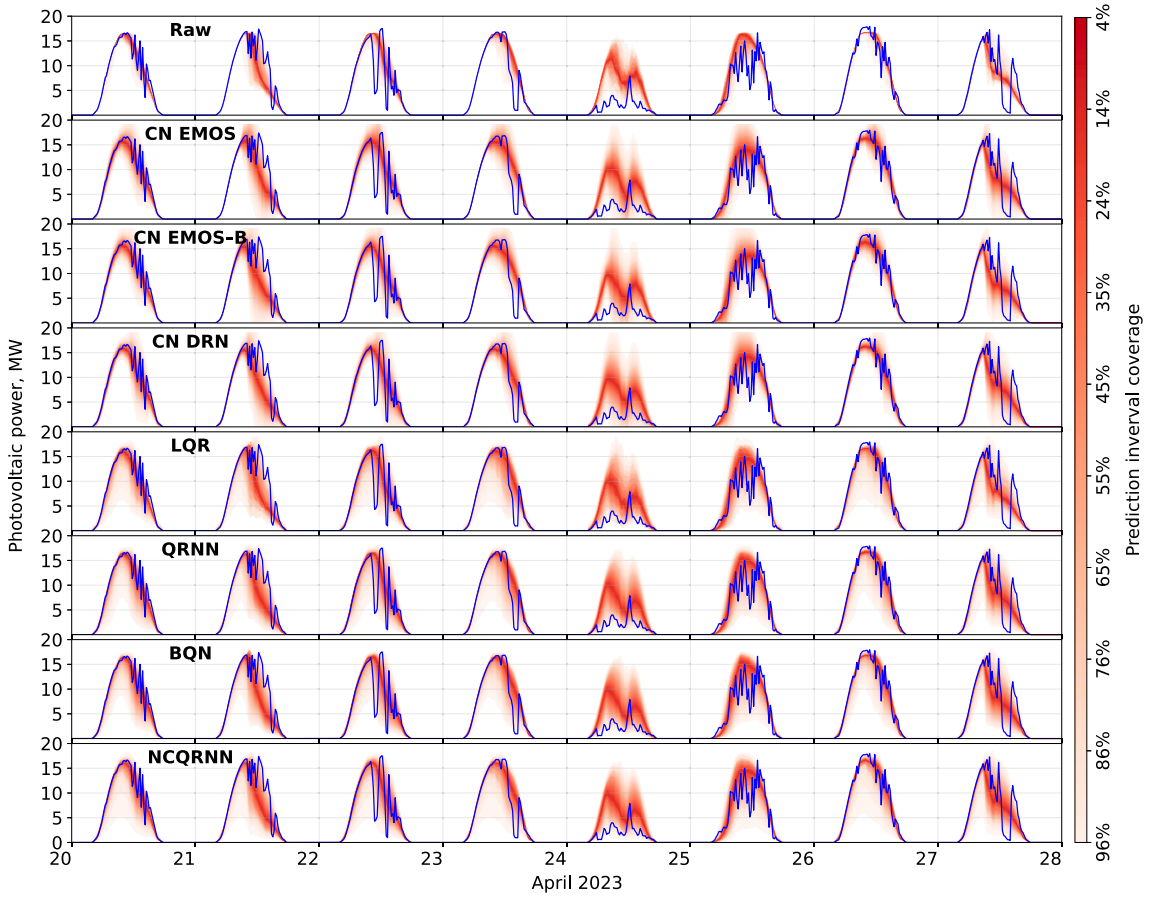


Fig. 10. Sample time series plot of the measured PV power production (blue) and the of the raw and calibrated ensemble probabilistic forecasts (red) for the Paks PV plant.

From the competing post-processing methods, the advanced machine learning-based non-linear non-parametric quantile regression models (QRNN, BQN, and NCQRNN) behave very similarly and consistently outperform the other four approaches. Among these three methods, the QRNN exhibits the best overall skill. In general, non-parametric methods show superior predictive performance compared to the parametric models, which matches the results of, for instance, [68]; nevertheless, the best-performing parametric CN DRN shows performance on par with the least skillful non-parametric LQR. Furthermore, our study also confirms that machine learning-based approaches surpass their traditional statistical counterparts (CN DRN vs. CN EMOS or QRNN vs. LQR), even though here the full potential of neural networks stemming from their capability of easily accommodating additional relevant covariates was not exploited; all models relied merely on the same set of inputs given by PV power ensemble forecasts or their summary statistics. The main advantage of the machine learning-based approaches in our study thus lies in their ability to flexibly model possibly non-linear relationships between the inputs and the various types of outputs characterizing predictive distributions.

6. Conclusions

We investigated the forecast skill of seven different models for statistical post-processing of PV power ensemble predictions using a wide range of evaluation metrics. The chosen pool of post-processing methods represents, on the one hand, both parametric (CN EMOS, CN EMOS-B, CN DRN) and non-parametric techniques (LQR, QRNN, BQN, NCQRNN), and on the other hand, both traditional statistical (CN EMOS, CN EMOS-B, LQR) and machine learning-based approaches (CN DRN, QRNN, BQN, NCQRNN). To the best of our knowledge, our study represents the first broad comparison of parametric and non-parametric post-processing methods for solar energy forecasting in a model chain approach. In line with the findings of, e.g., [9] or [23], our study confirms the superiority of any form of statistical post-processing over the raw PV power forecasts, with the non-parametric models displaying the best overall predictive performance.

The post-processing methods considered in our study provide several avenues for further improvements and analysis. For example, comparisons with alternative non-parametric approaches that are not directly based on quantile regression, such as isotonic distributional regression [69], member-by-member post-processing [70], or conformal prediction [71] might be of interest and could help to identify alternative approaches, although previous studies generally indicate a similar predictive performance to EMOS [16]. In addition, while our study indicates the superiority of non-parametric techniques over the parametric models, it would be interesting to investigate under what data availability conditions the parametric approaches become more efficient than, for instance, the quantile regression methods. Further, it has been noted in the post-processing literature that a key reason for the success of modern machine learning-based methods is their capability to include additional covariates, e.g., [17]. Therefore, the neural network-based parametric and non-parametric approaches might be further improved by extending their inputs with weather predictions from the NWP system. That said, [23] noted only minor improvements when doing so. Another route towards improving the predictive performance might be a more sophisticated use of the lead time information, as for example proposed by [72], or incorporating expert variables into the set of possible covariates, followed by a feature selection procedure, as discussed in [73]. Further,

instead of applying a single model chain only, it would also be possible to consider an ensemble of possible model chains [9], leading to a multi-model ensemble forecast of PV power production. Given the relevance of probabilistic energy forecasts in grid operations and electricity markets, the statistical evaluation based on scoring rules considered here should further be accompanied by considerations of other aspects, including the economic impacts of improved forecasts [5]. Our study was limited to PV plants in Hungary. Extensions to other geographical regions would be of interest, but are not straightforward due to limitations in the public availability of PV power production data.

In recent years, machine learning–based, purely data-driven weather models have advanced rapidly. Notable examples include Pangu-Weather [74], GraphCast [75], and AIFS [76], which provide deterministic forecasts, as well as more recent ensemble prediction systems such as GenCast [77] and AIFS-CRPS [78]. These models now surpass physics-based NWP approaches for a range of weather variables. A key question in the context of solar energy forecasting is whether predictions from these data-driven weather models might replace NWP ensemble forecasts, and which role post-processing methods could play. For example, recent research has demonstrated that data-driven and physics-based weather models might equally benefit from post-processing [79,80]. However, most of the current data-driven weather models do not provide relevant outputs such as GHI, although there have been significant recent developments including the FuXi-2.0 model [81], which explicitly targets solar and wind energy forecasting.

CRediT authorship contribution statement

Martin János Mayer: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Ágnes Baran:** Writing – original draft, Validation, Software, Methodology, Investigation, Data curation. **Sebastian Lerch:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Nina Horat:** Software, Methodology. **Dazhi Yang:** Writing – original draft, Conceptualization. **Sándor Baran:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. One of the authors is a Subject Editor for this journal and was not involved in the editorial review or the decision to publish this article.

Acknowledgments

The authors thank Norbert Péter from MVM Green Generation Ltd. for the PV plant design and production data and the HungaroMet Hungarian Meteorological Service for providing access to the ECMWF's Meteorological Archival and Retrieval System. Martin János Mayer was supported by the National Research, Development and Innovation Fund under the project no. OTKA-FK 142702, and the Hungarian Academy of Sciences through the János Bolyai Research Scholarship. Ágnes Baran and Sándor Baran were supported by the Hungarian National Research, Development and Innovation Office under Grant K142849. The work leading to this paper was done, in part, during the visit of Sándor Baran to the Heidelberg Institute for Theoretical Studies in July 2025 as a guest researcher. Sebastian Lerch and Nina Horat acknowledge funding from the Vector Stiftung within the Young Investigator Group "Artificial Intelligence for Probabilistic Weather Forecasting". Last, but not least, the authors thank the three anonymous reviewers, whose constructive comments helped improve the manuscript.

Appendix A. Significance of score differences

To address the significance of differences between the various forecasts in terms of the mean CRPS and MAE, we consider two different approaches. On the one hand, we complement the mean scores and some of the skill scores with Gaussian 95% confidence intervals using standard deviations based on 2000 stationary block bootstrap samples with random block lengths drawn from a geometric distribution [82]. On the other hand, for each location and observation time, we perform pairwise Diebold-Mariano (DM) [83] tests for equal predictive performance and report the proportion of cases where the difference in mean CRPS and MAE is significant at a 5% level. Following the suggestions of [84], to control the false discovery rate in simultaneous testing, we apply the Benjamini-Hochberg algorithm [85]. To avoid the distortion resulting from very low observed and forecasted PV production values, the following analysis restricts the time interval of observations to the hours of peak PV power production between 06:00 and 16:00 UTC.

According to Fig. A.11(a), the parametric approaches and the simple LQR are significantly behind the QRNN in terms of the CRPSS to almost all considered observation times, and the same applies for the other two advanced nonparametric methods (BQN and NCQRNN, not shown). Considering the MAES of the median (Fig. A.11(b), the situation slightly changes, as the skill scores of the CN DRN approach are mainly between the lower and upper confidence bounds for the MAES of the QRNN (and the BQN and NCQRNN as well, not shown). Finally, even the CRPSS values of the worst performing CN EMOS approach are significantly positive during the whole observation period (not shown), the minimal value of the 95% lower bound is 7.27%, and the MAES of this parametric method is not significantly positive at a 5% level only at 06:00, 06:30, 06:45 and 14:30, 14:45, 15:00 UTC (not shown).

Furthermore, Tables A.6 and A.7 confirm that compared to the raw ensemble, any form of post-processing significantly and consistently improves the mean CRPS of the probabilistic predictions and the MAE of the median forecasts. They also verify that where Tables 5 and 1 show substantial differences in terms of the mean CRPS and MAE between the QRNN, BQN, and NCQRNN approaches and the other four post-processing methods, these differences are significant at a 5% level.

Finally, Fig. A.12 approaches the question of significance in the score differences from another angle. Each entry summarizes the results of 287 parallel pairwise one-sided DM tests (7 locations, 41 observation times) by reporting the proportion of cases where the difference in predictive performance of the compared forecasts is significant at a 5% level. The results of these DM tests are completely in line with the confidence interval-based findings: the differences between QRNN, BQN, and NCQRNN are minor, the raw forecast is significantly behind the post-processed ones in more than 98% of the cases, and the differences between the various post-processing approaches in terms of the MAE of the median are less pronounced than in terms of the mean CRPS.

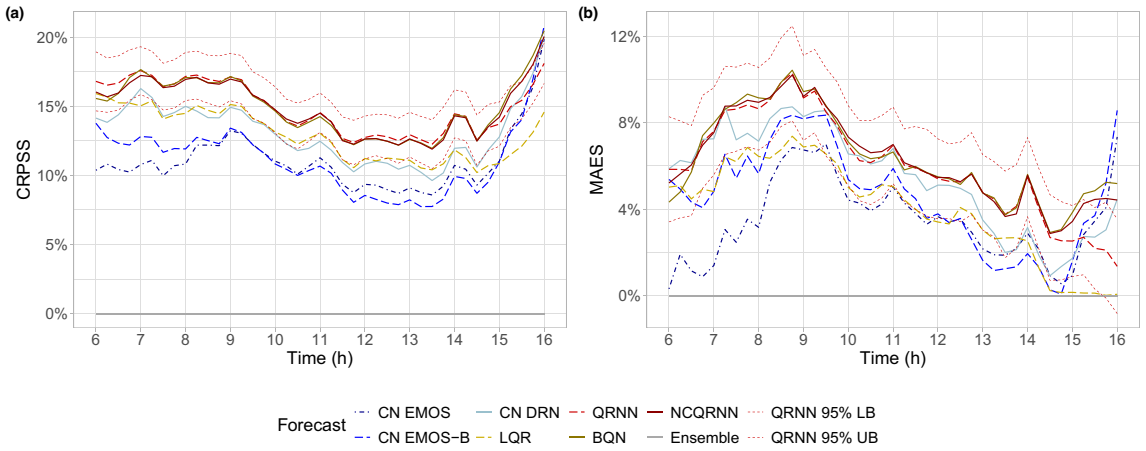


Fig. A.11. CRPSS (a) and MAES of the median (b) of post-processed PV power forecasts normalized to the mean daytime power production of the PV plants with respect to the raw ensemble as functions of the observation time (06:00–16:00 UTC) together with 95% confidence intervals for the QRNN method.

Table A.6
Mean CRPS of post-processed and raw PV power forecasts normalized to the mean daytime power production of the PV plants for the 06:00–16:00 UTC time period together with 95% confidence intervals.

Forecast	Overall	Bodajk	Cegléd	Felsőzsolca
CN EMOS	22.41±0.23%	23.13±0.63%	22.37±0.53%	24.77±0.71%
CN EMOS–B	22.44±0.23%	23.10±0.64%	22.26±0.53%	24.84±0.72%
CN DRN	21.92±0.23%	22.36±0.62%	21.79±0.55%	24.15±0.74%
LQR	21.90±0.23%	22.57±0.61%	21.64±0.53%	24.21±0.69%
QRNN	21.42±0.23%	21.99±0.61%	21.23±0.54%	23.49±0.72%
BQN	21.44±0.23%	22.00±0.61%	21.22±0.55%	23.46±0.73%
NCQRNN	21.44±0.23%	22.03±0.61%	21.21±0.54%	23.59±0.72%
Ensemble	25.08±0.26%	26.09±0.71%	24.83±0.61%	27.58±0.78%
	Fertőszéplak	Magyarsarlós	Paks	Újkígyós
CN EMOS	23.62±0.69%	20.47±0.54%	20.96±0.59%	21.63±0.57%
CN EMOS–B	23.81±0.69%	20.41±0.54%	20.94±0.58%	21.80±0.57%
CN DRN	23.14±0.69%	20.19±0.54%	20.52±0.59%	21.35±0.59%
LQR	23.16±0.66%	20.28±0.53%	20.35±0.58%	21.15±0.56%
QRNN	22.57±0.68%	19.95±0.54%	20.04±0.57%	20.71±0.57%
BQN	22.63±0.69%	20.00±0.53%	20.02±0.58%	20.81±0.58%
NCQRNN	22.67±0.68%	19.85±0.54%	20.04±0.58%	20.75±0.58%
Ensemble	25.93±0.79%	24.66±0.61%	23.08±0.65%	23.45±0.64%

Table A.7
MAE of the median of post-processed and raw PV power forecasts normalized to the mean daytime power production of the PV plants for the 06:00 – 16:00 UTC time period together with 95% confidence intervals.

Forecast	Overall	Bodajk	Cegléd	Felsőzsolca
CN EMOS	31.31±0.33%	32.32±0.90%	31.58±0.78%	34.51±1.00%
CN EMOS–B	31.03±0.32%	31.94±0.90%	31.06±0.76%	34.34±0.99%
CN DRN	30.69±0.33%	31.49±0.89%	30.77±0.79%	33.81±1.03%
LQR	31.11±0.33%	32.10±0.90%	30.86±0.80%	34.32±1.01%
QRNN	30.46±0.34%	31.38±0.88%	30.25±0.81%	33.23±1.06%
BQN	30.40±0.34%	31.26±0.88%	30.17±0.82%	33.04±1.07%
NCQRNN	30.40±0.34%	31.32±0.89%	30.17±0.81%	33.28±1.05%
Ensemble	32.53±0.34%	33.89±0.92%	32.20±0.83%	35.57±1.02%
	Fertőszéplak	Magyarsarlós	Paks	Újkígyós
CN EMOS	33.12±0.95%	28.12±0.78%	29.31±0.85%	30.29±0.84%
CN EMOS–B	32.98±0.95%	27.85±0.76%	28.95±0.81%	30.20±0.83%
CN DRN	32.47±0.96%	27.89±0.78%	28.75±0.85%	29.78±0.84%
LQR	32.97±0.99%	29.11±0.78%	28.77±0.85%	29.77±0.86%
QRNN	32.15±1.00%	28.50±0.80%	28.47±0.84%	29.36±0.87%
BQN	32.09±0.99%	28.65±0.80%	28.37±0.85%	29.33±0.86%
NCQRNN	32.11±0.99%	28.36±0.80%	28.41±0.85%	29.26±0.86%
Ensemble	33.56±1.01%	32.05±0.81%	30.10±0.88%	30.43±0.85%

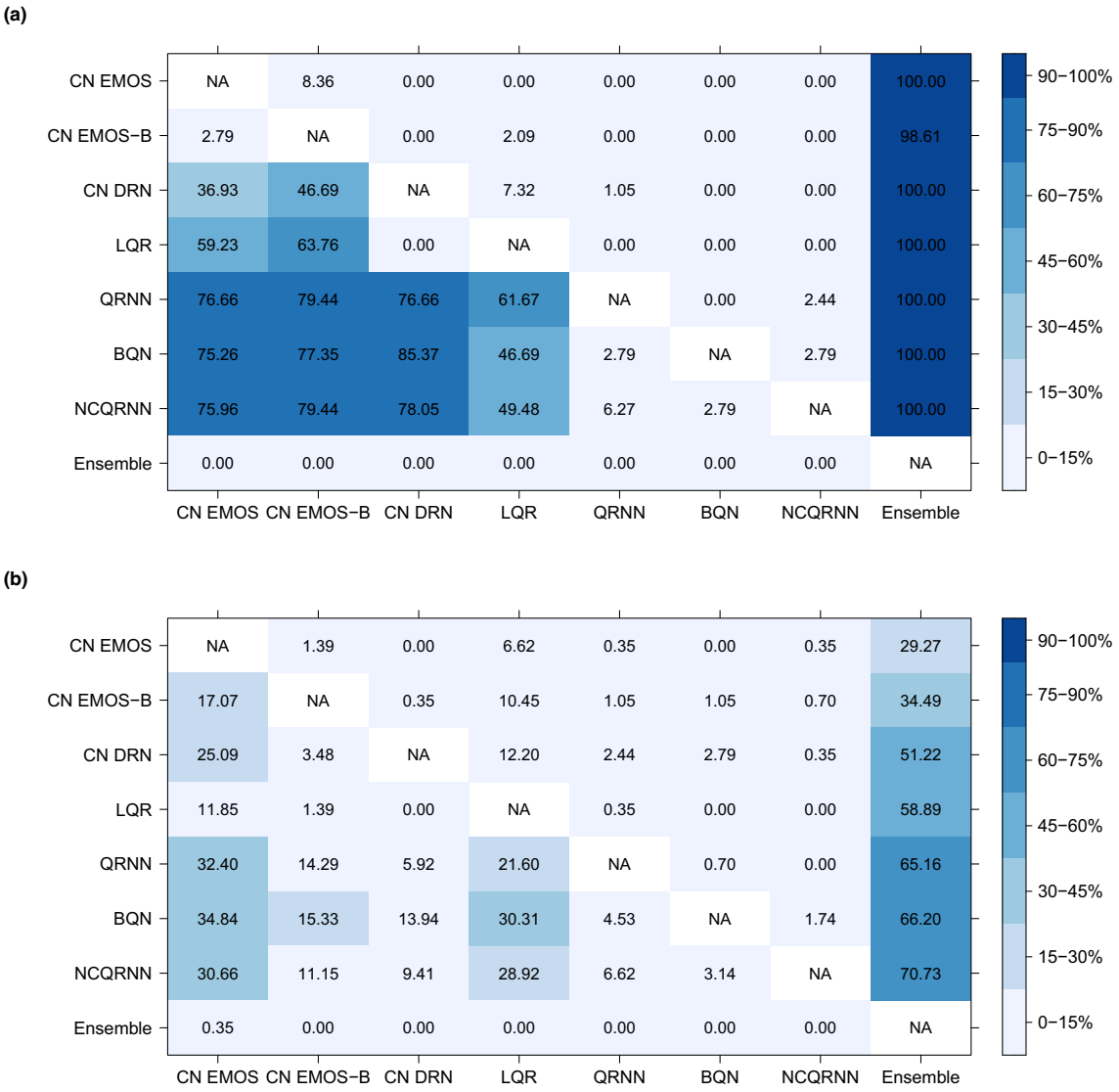


Fig. A.12. Proportion of cases where the null hypothesis of equal predictive performance in terms of the mean CRPS (a) and MAE (b) of the corresponding one-sided DM test is rejected at a 5% level of significance in favor of the forecast in the row when compared with the forecast in the column.

Data availability

The data that has been used in the present study is confidential.

References

[1] T. Hong, P. Pinson, Y. Wang, R. Weron, D. Yang, H. Zareipour, Energy forecasting: a review and outlook, IEEE Open Access J. Power Energy 7 (2020) 376–388, https://doi.org/10.1109/OAJPE.2020.3029979

[2] D. Yang, W. Wang, C.A. Gueymard, T. Hong, J. Kleissl, J. Huang, M.J. Perez, R. Perez, J.M. Bright, X. Xia, D. van der Meer, I.M. Peters, A review of solar forecasting, its dependence on atmospheric sciences and implications for grid integration: towards carbon neutrality, Renew. Sustain. Energy Rev. 161 (2022a) 112348, https://doi.org/10.1016/j.rser.2022.112348

[3] D. Yang, W. Wang, X. Xia, A concise overview on solar resource assessment and forecasting, Adv. Atmos. Sci. 39 (2022b) 1239–1251, https://doi.org/10.1007/s00376-021-1372-8

[4] D. Yang, D. van der Meer, Post-processing in solar forecasting: ten overarching thinking tools, Renew. Sustain. Energy Rev. 140 (2021) 110735, https://doi.org/10.1016/j.rser.2021.110735

[5] T. Gneiting, S. Lerch, B. Schulz, Probabilistic solar forecasting: benchmarks, post-processing, verification, Sol. Energy 252 (2023) 72–80, https://doi.org/10.1016/j.solener.2022.12.054

[6] P. Lauret, M. David, P. Pinson, Verification of solar irradiance probabilistic forecasts, Sol. Energy 194 (2019) 254–271, https://doi.org/10.1016/j.solener.2019.10.041

[7] T. Gneiting, D. Wolffram, J. Resin, K. Kraus, J. Bracher, T. Dimitriadis, V. Hagenmeyer, A.I. Jordan, S. Lerch, K. Phipps, et al., Model diagnostics and forecast evaluation for quantiles, Annu. Rev. Stat. Its Appl. 10 (2023) 597–621, https://doi.org/10.1146/annurev-statistics-032921-020240

[8] J.J. Roberts, A.A. Mendiburu Zevallos, A.M. Cassula, Assessment of photovoltaic performance models for system simulation, Renew. Sustain. Energy Rev. 72 (2017) 1104–1123, https://doi.org/10.1016/j.rser.2016.10.022

[9] M.J. Mayer, D. Yang, Probabilistic photovoltaic power forecasting using a calibrated ensemble of model chains, Renew. Sustain. Energy Rev. 168 (2022) 112821, https://doi.org/10.1016/j.rser.2022.112821

[10] D. Yang, X. Xia, M.J. Mayer, A tutorial review of the solar power curve: regressions, model chains, and their hybridization and probabilistic extensions, Adv. Atmos. Sci. 41 (2024) 1023–1067, https://doi.org/10.1007/s00376-024-3229-4

[11] R. Buizza, Ensemble forecasting and the need for calibration, In S. Vannitsem, D.S. Wilks, J.W. Messner (Eds.), Statistical Postprocessing of Ensemble Forecasts, Elsevier, Amsterdam, 2018, pp. 15–48, https://doi.org/10.1016/B978-0-12-812372-0.00002-9

[12] S. Vannitsem, J.B. Bremnes, J. Demaeyer, G.R. Evans, J. Flowerdew, S. Hemri, S. Lerch, N. Roberts, S. Theis, A. Atencia, Z. Ben Bouallègue, J. Bhend, M. Dabernig, L. De Cruz, L. Hieta, O. Mestre, L. Moret, I.O. Plenković, M. Schmeits, M. Taillardat, J. Van den Bergh, B. Van Schaeybroeck, K. Whan, J. Ylhaisi, Statistical postprocessing for weather forecasts – review, challenges and avenues in a big data world, Bull. Am. Meteorol. Soc. 102 (2021) E681–E699, https://doi.org/10.1175/BAMS-D-19-0308.1

[13] M. Taillardat, O. Mestre, M. Zamo, P. Naveau, Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics, Mon. Weather Rev. 144 (2016) 2375–2393, https://doi.org/10.1175/MWR-D-15-0260.1

[14] S. Rasp, S. Lerch, Neural networks for postprocessing ensemble weather forecasts, Mon. Weather Rev. 146 (2018) 3885–3900, https://doi.org/10.1175/MWR-D-18-0187.1

[15] S.E. Haupt, W. Chapman, S.V. Adams, C. Kirkwood, J.S. Hosking, N.H. Robinson, S. Lerch, A.C. Subramanian, Towards implementing artificial intelligence post-processing in weather and climate: proposed actions from the Oxford 2019 workshop, Philos. Trans. R. Soc. A: Math. Phys. Eng. Sci. 379 (2021) 20200091, https://doi.org/10.1098/rsta.2020.0091

[16] B. Schulz, S. Lerch, Machine learning methods for postprocessing ensemble forecasts of wind gusts: a systematic comparison, Mon. Weather Rev. 150 (2022) 235–257, https://doi.org/10.1175/MWR-D-21-0150.1

[17] J. Demaeyer, J. Bhend, S. Lerch, C. Primo, B. Van Schaeybroeck, A. Atencia, Z. Ben Bouallègue, J. Chen, M. Dabernig, G. Evans, J. Faganeli Pucer, B. Hooper, N. Horat, D. Jobst, J. Merše, P. Mlakar, A. Möller, O. Mestre, M. Taillardat, S. Vannitsem, The EUPPBench postprocessing benchmark dataset v1.0, Earth Syst. Sci. Data 15 (2023) 2635–2653, https://doi.org/10.5194/essd-15-2635-2023

[18] K. Bakker, K. Whan, W. Knap, M. Schmeits, Comparison of statistical post-processing methods for probabilistic NWP forecasts of solar radiation, Sol. Energy 191 (2019) 138–150, https://doi.org/10.1016/j.solener.2019.08.044

[19] J. Le Gal La Salle, J. Badosa, M. David, P. Pinson, P. Lauret, Added-value of ensemble prediction system on the quality of solar irradiance probabilistic forecasts, Renew. Energy 162 (2020) 1321–1339, https://doi.org/10.1016/j.renene.2020.07.042

[20] B. Schulz, M. El Ayari, S. Lerch, S. Baran, Post-processing numerical weather prediction ensembles for probabilistic solar irradiance forecasting, Sol. Energy 220 (2021) 1016–1031, https://doi.org/10.1016/j.solener.2021.03.023

[21] Á. Baran, S. Baran, A two-step machine learning approach to statistical post-processing of weather forecasts for power generation, Q. J. R. Meteorol. Soc. 150 (2024) 1029–1047, https://doi.org/10.1002/qj.4635

[22] M. Song, D. Yang, S. Lerch, X. Xia, G.M. Yagli, J.M. Bright, Y. Shen, B. Liu, X. Liu, M.J. Mayer, Non-crossing quantile regression neural network as a calibration tool for ensemble weather forecasts, Adv. Atmos. Sci. 41 (2024) 1417–1437, https://doi.org/10.1007/s00376-023-3184-5

[23] N. Horat, S. Klerings, S. Lerch, Improving model chain approaches for probabilistic solar energy forecasting through post-processing and machine learning, Adv. Atmos. Sci. 42 (2025) 297–312, https://doi.org/10.1007/s00376-024-4219-2

[24] W. Wang, D. Yang, T. Hong, J. Kleissl, An archived dataset from the ECMWF ensemble prediction system for probabilistic solar power forecasting, Sol. Energy 248 (2022) 64–75, https://doi.org/10.1016/j.solener.2022.10.062

[25] S. Theocharides, G. Makrides, A. Livera, M. Theristis, P. Kaimakis, G.E. Georghiou, Day-ahead photovoltaic power production forecasting methodology based on machine learning and statistical post-processing, Appl. Energy 268 (2020) 115023, https://doi.org/10.1016/j.apenergy.2020.115023

[26] M.J. Mayer, D. Yang, Optimal place to apply post-processing in the deterministic photovoltaic power forecasting workflow, Appl. Energy 371 (2024) 123681, https://doi.org/10.1016/j.apenergy.2024.123681

[27] K. Phipps, S. Lerch, M. Andersson, R. Mikut, V. Hagenmeyer, N. Ludwig, Evaluating ensemble post-processing for wind power forecasts, Wind Energy 25 (2022) 1379–1405, https://doi.org/10.1002/we.2736

[28] T. Gneiting, A.E. Raftery, A.H. Westveld, T. Goldman, Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation, Mon. Weather Rev. 133 (2005) 1098–1118, https://doi.org/10.1175/MWR2904.1

[29] J.W. Messner, G.J. Mayr, A. Zeileis, Nonhomogeneous boosting for predictor selection in ensemble postprocessing, Mon. Weather Rev. 145 (2017) 137–147, https://doi.org/10.1175/MWR-D-16-0088.1

[30] J.W. Taylor, A quantile regression neural network approach to estimating the conditional density of multiperiod returns, J. Forecast. 19 (2000) 299–311, 10.1002/1099-131X(200007)19:4<299::AID-FOR775>3.0.CO;2-V

[31] J.B. Bremnes, Ensemble postprocessing using quantile function regression based on neural networks and bernstein polynomials, Mon. Weather Rev. 148 (2020) 403–414, https://doi.org/10.1175/MWR-D-19-0227.1

[32] H.E. Beck, T.R. McVicar, N. Vergopolan, A. Berg, N.J. Lutsko, A. Dufour, Z. Zeng, X. Jiang, A.I.J.M. van Dijk, D.G. Miralles, High-resolution (1 km) köppen-geiger maps for 1901–2099 based on constrained cmip6 projections, Sci. Data 10 (2023) 724, https://doi.org/10.1038/s41597-023-02549-6

[33] M. Lefèvre, A. Oumbe, P. Blanc, B. Espinar, B. Gschwind, Z. Qu, L. Wald, M. Schroedter-Homscheidt, C. Hoyer-Klick, A. Arola, A. Benedetti, J.W. Kaiser, J.J. Morcrette, McClear: a new model estimating downwelling solar radiation at ground level in clear-sky conditions, Atmos. Meas. Tech. 6 (2013) 2403–2418, https://doi.org/10.5194/AMT-6-2403-2013

[34] M.J. Mayer, G. Gróf, Techno-economic optimization of grid-connected, ground-mounted photovoltaic power plants by genetic algorithm based on a comprehensive mathematical model, Solar Energy 202 (2020) 210–226, https://doi.org/10.1016/j.solener.2020.03.109

[35] M.J. Mayer, G. Gróf, Extensive comparison of physical models for photovoltaic power forecasting, Appl. Energy 283 (2021) 116239, https://doi.org/10.1016/j.apenergy.2020.116239

[36] I. Reda, A. Andreas, Solar position algorithm for solar radiation applications, Sol. Energy 76 (2004) 577–589, https://doi.org/10.1016/j.solener.2003.12.003

[37] D. Yang, Temporal-resolution cascade model for separation of 1-min beam and diffuse irradiance, J. Renew. Sustain. Energy 13 (2021) 056101, https://doi.org/10.1063/5.0067997

[38] D. Yang, Estimating 1-min beam and diffuse irradiance from the global irradiance: a review and an extensive worldwide comparison of latest separation models at 126 stations, Renew. Sustain. Energy Rev. 159 (2022) 112195, https://doi.org/10.1016/j.rser.2022.112195

[39] D. Yang, Y. Gu, M.J. Mayer, C.A. Gueymard, W. Wang, J. Kleissl, M. Li, Y. Chu, J.M. Bright, Regime-dependent 1-min irradiance separation model with Climatology clustering, Renew. Sustain. Energy Rev. 189 (2024) 113992, https://doi.org/10.1016/j.rser.2023.113992

[40] R. Perez, P. Ineichen, R. Seals, J. Michalsky, R. Stewart, Modeling daylight availability and irradiance components from direct and global irradiance, Sol. Energy 44 (1990) 271–289, https://doi.org/10.1016/0038-092X(90)90055-H

[41] D. Yang, Solar radiation on inclined surfaces: corrections and benchmarks, Solar Energy 136 (2016) 288–302, https://doi.org/10.1016/j.solener.2016.06.062

[42] N. Martin, J.M. Ruiz, Calculation of the PV modules angular losses under field conditions by means of an analytical model, Sol. Energy Mater. Sol. Cells 70 (2001) 25–38, https://doi.org/10.1016/S0927-0248(00)00408-6

[43] M.J. Mayer, Impact of the tilt angle, inverter sizing factor and row spacing on the photovoltaic power forecast accuracy, Appl. Energy 323 (2022) 119598, https://doi.org/10.1016/j.apenergy.2022.119598

[44] M. Mattei, G. Notton, C. Cristofari, M. Muselli, P. Poggi, Calculation of the polycrystalline PV module temperature using a simple method of energy balance, Renew. Energy 31 (2006) 553–567, https://doi.org/10.1016/j.renene.2005.03.010

[45] W. De Soto, S.A. Klein, W.A. Beckman, Improvement and validation of a model for photovoltaic array performance, Solar Energy 80 (2006) 78–88, https://doi.org/10.1016/j.solener.2005.06.010

[46] A. Driesse, P. Jain, S. Harrison, Beyond the curves: modeling the electrical efficiency of photovoltaic inverters, in: 2008 33rd IEEE Photovolatic Specialists Conference, IEEE, 2008, pp. 1–6, https://doi.org/10.1109/PVSC.2008.4922827

[47] T.L. Thorarinsdottir, T. Gneiting, Probabilistic forecasts of wind speed: ensemble model output statistics by using heteroscedastic censored regression, J. R. Stat. Soc. Ser. A Stat. Soc. 173A (2010) 371–388, https://doi.org/10.1111/j.1467-985X.2009.00616.x

[48] S. Baran, S. Lerch, Log-normal distribution based EMOS models for probabilistic wind speed forecasting, Q. J. R. Meteorol. Soc. 141 (2015) 2289–2299, https://doi.org/10.1002/qj.2521

[49] M. Scheuerer, Probabilistic quantitative precipitation forecasting using ensemble model output statistics, Q. J. R. Meteorol. Soc. 140 (2014) 1086–1096, https://doi.org/10.1002/qj.2183

[50] S. Baran, D. Nemoda, Censored and shifted gamma distribution based EMOS model for probabilistic quantitative precipitation forecasting, Environmetrics 27 (2016) 280–292, https://doi.org/10.1002/env.2391

[51] T. Gneiting, A.E. Raftery, Strictly proper scoring rules, prediction and estimation, J. Am. Stat. Assoc. 102 (2007) 359–378, https://doi.org/10.1198/016214506000001437

[52] D.S. Wilks ed), Statistical Methods in the Atmospheric Sciences, fourth ed, Elsevier, Amsterdam, 2019, https://doi.org/10.1016/C2017-0-03921-6

[53] J.W. Messner, G.J. Mayr, A. Zeileis, Heteroscedastic censored and truncated regression with crch, The R J. 8 (2016) 173–181, https://doi.org/10.32614/RJ-2016-012

[54] A.J. Cannon, Quantile regression neural networks: implementation in r and application to precipitation downscaling, Comput. Geosci. 37 (2011) 1277–1284, https://doi.org/10.1016/j.cageo.2010.07.005

[55] R. Koenker, Quantile Regression, Cambridge University Press, Cambridge, UK, 2005, https://doi.org/10.1017/CBO9780511754098

[56] T. Gneiting, Making and evaluating point forecasts, J. Am. Stat. Assoc. 106 (2011) 746–762, https://doi.org/10.1198/jasa.2011.r10138

[57] S. Kolassa, Why the "best" point forecast depends on the error or accuracy measure, Int. J. Forecast. 36 (2020) 208–211, https://doi.org/10.1016/j.ijforecast.2019.02.017

[58] M.J. Mayer, D. Yang, Calibration of deterministic NWP forecasts and its impact on verification, Int. J. Forecast. 39 (2023) 981–991, https://doi.org/10.1016/j.ijforecast.2022.03.008

[59] A. Jordan, F. Krüger, S. Lerch, Evaluating probabilistic forecasts with scoringrules, J. Stat. Softw. 90 (2019) 1–37, https://doi.org/10.18637/jss.v090.i12

[60] F. Krüger, S. Lerch, T. Thorarinsdottir, T. Gneiting, Predictive inference based on markov chain monte carlo output, Int. Stat. Rev. 89 (2021) 274–301, https://doi.org/10.1111/insr.12405

[61] H. Hersbach, Decomposition of the continuous ranked probability score for ensemble prediction systems, Weather Forecast. 15 (2000) 559–570, https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2

[62] J. Bröcker, Reliability, sufficiency, and the decomposition of proper scores, Q. J. R. Meteorol. Soc. 135 (2009) 1512–1519, https://doi.org/10.1002/qj.456

[63] T. Gneiting, R. Ranjan, Comparing density forecasts using thresholdand quantile-weighted scoring rules, J. Bus. Econ. Stat. 29 (2011) 411–422, https://doi.org/10.1198/jbes.2010.08110

[64] L. Delle Monache, J.P. Hacker, Y. Zhou, X. Deng, R.B. Stull, Probabilistic aspects of meteorological and ozone regional ensemble forecasts, J. Geophys. Res.: Atmos. 111 (2006) D24307, https://doi.org/10.1029/2005JD006917

[65] B. Schulz, L. Köhler, S. Lerch, Aggregating distribution forecasts from deep ensembles, 2022. Preprint, available at https://arxiv.org/abs/2204.02291.

[66] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: a next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19, New York, NY, USA, Association for Computing Machinery, 2019, pp. 2623–2631, https://doi.org/10.1145/3292500.3330701

[67] J. Bröcker, Evaluating raw ensembles with the continuous ranked probability score, Q. J. R. Meteorol. Soc. 138 (2012) 1611–1617, https://doi.org/10.1002/qj.1891

[68] S. Baran, J.C. Marín, O. Cuevas, M. Díaz, M. Szabó, O. Nicolis, M. Lakatos, Machine-learning-based probabilistic forecasting of solar irradiance in Chile, Adv. Stat.

Climatol. Meteorol. Oceanogr. 11 (2025) 89–105, https://doi.org/10.5194/ascmo-11-89-2025
[69] A. Henzi, J.F. Ziegel, T. Gneiting, Isotonic distributional regression, J. R. Stat. Soc. Ser. B Stat. Methodol. 83 (2021) 963–993, https://doi.org/10.1111/rssb.12450
[70] B. Van Schaeybroeck, S. Vannitsem, Ensemble post-processing using member-by-member approaches: theoretical aspects, Q. J. R. Meteorol. Soc. 141 (2015) 807–818, https://doi.org/10.1002/qj.2397
[71] Y. Renkema, L. Visser, T. AlSkaif, Enhancing the reliability of probabilistic PV power forecasts using conformal prediction, Sol. Energy Adv. 4 (2024) 100059, https://doi.org/10.1016/j.seja.2024.100059
[72] P. Mlakar, J. Merše, J. Faganeli Pucer, Ensemble weather forecast post-processing with a flexible probabilistic neural network approach, Q. J. R. Meteorol. Soc. 150 (2024) 4156–4177, https://doi.org/10.1002/qj.4809
[73] L. Visser, T. AlSkaif, J. Hu, A. Louwen, W. van Sark, On the value of expert knowledge in estimation and forecasting of solar photovoltaic power generation, Solar Energy 251 (2023) 86–105, https://doi.org/10.1016/j.solener.2023.01.019
[74] K. Bi, L. Xie, H. Zhang, X. Chen, X. Gu, Q. Tian, Accurate medium-range global weather forecasting with 3d neural networks, Nature 619 (2023) 533–538, https://doi.org/10.1038/s41586-023-06185-3
[75] R. Lam, A. Sanchez-Gonzalez, M. Willson, P. Wirnsberger, M. Fortunato, F. Alet, S. Ravuri, T. Ewalds, Z. Eaton-Rosen, W. Hu, A. Merose, S. Hoyer, G. Holland, O. Vinyals, J. Stott, A. Pritzel, S. Mohamed, P. Battaglia, Learning skillful medium-range global weather forecasting, Science 382 (2023) 1416–1421, https://doi.org/10.1126/science.adi2336
[76] S. Lang, M. Alexe, M. Chantry, J. Dramsch, F. Pinault, B. Raoult, M.C.A. Clare, C. Lessig, M. Maier-Gerber, L. Magnusson, Z.B. Bouallègue, A.P. Nemesio, P.D. Dueben, A. Brown, F. Pappenberger, F. Rabier, AIFS – ECMWF's data-driven forecasting system, 2024. Preprint, available at https://arxiv.org/abs/2406.01465.
[77] I. Price, A. Sanchez-Gonzalez, F. Alet, T.R. Andersson, A. El-Kadi, D. Masters, T. Ewalds, J. Stott, S. Mohamed, P. Battaglia, R. Lam, M. Willson, Probabilistic weather forecasting with machine learning, Nature 637 (2025) 84–90, https://doi.org/10.1038/s41586-024-08252-9
[78] S. Lang, M. Alexe, M.C.A. Clare, C. Roberts, R. Adewoyin, Z.B. Bouallègue, M. Chantry, J. Dramsch, P.D. Dueben, S. Hahner, P. Maciel, A. Prieto-Nemesio, C. O'Brien, F. Pinault, J. Polster, B. Raoult, S. Tietsche, M. Leutbecher, AIFS-CRPS: ensemble forecasting using a model trained with a loss function based on the continuous ranked probability score, 2024. Preprint, available at https://arxiv.org/abs/2412.15832.
[79] J.B. Bremnes, T.N. Nipen, I.A. Seierstad, Evaluation of forecasts by a global data-driven weather model with and without probabilistic post-processing at Norwegian stations, Nonlinear Process. Geophys. 31 (2024) 247–257, https://doi.org/10.5194/npg-31-247-2024
[80] C. Bülte, N. Horat, J. Quinting, S. Lerch, Uncertainty quantification for data-driven weather models, Artif. Intell. Earth Syst. (2025), https://doi.org/10.1175/AIES-D-24-0049.1
[81] X. Zhong, L. Chen, X. Fan, W. Qian, J. Liu, H. Li, Fuxi-2.0: advancing machine learning weather forecasting model for practical applications, 2024. Preprint, available at https://arxiv.org/abs/2409.07188.
[82] D.N. Politis, J.P. Romano, The stationary bootstrap, J. Am. Stat. Assoc. 89 (1994) 1303–1313, https://doi.org/10.1080/01621459.1994.10476870
[83] F.X. Diebold, R.S. Mariano, Comparing predictive accuracy, J. Bus. Econ. Stat. 13 (1995) 253–263, https://doi.org/10.1080/07350015.1995.10524599
[84] D.S. Wilks, "the stippling shows statistically significant grid points": how research results are routinely overstated and overinterpreted, and what to do about it, Bull. Am. Meteorol. Soc. 97 (2016) 2263–2273, https://doi.org/10.1175/BAMS-D-15-00267.1
[85] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc. Ser. B 57 (1995) 289–300, https://doi.org/10.1111/j.2517-6161.1995.tb02031.x