ADVANCES IN ATMOSPHERIC SCIENCES, VOL. 42, FEBRUARY 2025, 297–312

• **Original Paper** •

Improving Model Chain Approaches for Probabilistic Solar Energy Forecasting through Post-processing and Machine Learning※

Nina HORAT[1], Sina KLERINGS[1], and Sebastian LERCH*[1,2]

[1]*Institute of Statistics*, *Karlsruhe Institute of Technology*, *Karlsruhe* 76185, *Germany*

[2]*Heidelberg Institute for Theoretical Studies*, *Heidelberg* 69118, *Germany*

(Received 6 June 2024; revised 19 September 2024; accepted 23 September 2024)

ABSTRACT

Weather forecasts from numerical weather prediction models play a central role in solar energy forecasting, where a cascade of physics-based models is used in a model chain approach to convert forecasts of solar irradiance to solar power production. Ensemble simulations from such weather models aim to quantify uncertainty in the future development of the weather, and can be used to propagate this uncertainty through the model chain to generate probabilistic solar energy predictions. However, ensemble prediction systems are known to exhibit systematic errors, and thus require post-processing to obtain accurate and reliable probabilistic forecasts. The overarching aim of our study is to systematically evaluate different strategies to apply post-processing in model chain approaches with a specific focus on solar energy: not applying any post-processing at all; post-processing only the irradiance predictions before the conversion; post-processing only the solar power predictions obtained from the model chain; or applying post-processing in both steps. In a case study based on a benchmark dataset for the Jacumba solar plant in the U.S., we develop statistical and machine learning methods for post-processing ensemble predictions of global horizontal irradiance (GHI) and solar power generation. Further, we propose a neural-network-based model for direct solar power forecasting that bypasses the model chain. Our results indicate that post-processing substantially improves the solar power generation forecasts, in particular when post-processing is applied to the power predictions. The machine learning methods for post-processing slightly outperform the statistical methods, and the direct forecasting approach performs comparably to the post-processing strategies.

Key words: solar forecasting, post-processing, probabilistic forecasting, machine learning, model chain

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

Article Highlights:

- Post-processing substantially improves solar power forecasts, particularly, when the post-processing is applied to the power predictions.
- Whether or not the GHI forecasts are post-processed before using them as input to the model chain plays an almost negligible role.
- Post-processing methods for GHI and photovoltaic power should make use of the hour of the day, either as a predictor or by utilizing separate models.
- A neural-network-based, direct forecasting model that bypasses the model chain performs comparably to the best post-processing strategy.

1. Introduction

Reducing greenhouse gas emissions and mitigating climate change requires a rapid transition towards renewable energy (Van der Meer et al., 2018). In addition to wind energy, photovoltaic (PV) solar power plays a pivotal role, with decreasing prices and increasing installed capacity in numerous countries. For example, PV power covered 12% of the gross electricity consumption in Germany on average in 2023, and temporarily more than two thirds of the electricity demand on sunny days (Fraunhofer Institute for Solar Energy Systems, 2024). In light of the volatile nature of renewable energy generation and its increasing importance, accurate and reliable forecasts of power generation from those sources are paramount for managing the electrical

※ This paper is a contribution to the special topic on Solar Energy Meteorology.

* Corresponding author: Sebastian LERCH
Email: sebastian.lerch@kit.edu

grid and to balance demand and supply (Gottwalt et al., 2017; Appino et al., 2018). A key development in the energy forecasting literature over recent years has been the transition from single-valued deterministic to probabilistic forecasts (Gneiting and Katzfuss, 2014; Haupt et al., 2019; Yang, 2019; Yang and van der Meer, 2021; Gneiting et al., 2023a), which allow for uncertainty quantification and can be issued in the form of probability distributions, quantiles, or prediction intervals (Lauret et al., 2019; Gneiting et al., 2023b).

Evidently, weather forecasts from numerical weather prediction (NWP) models are among the most important inputs to models for PV power forecasting. A widely used strategy is the conversion of global horizontal irradiance (GHI) forecasts from an NWP system to PV power forecasts via a model chain, potentially using predictions of other meteorological variables as additional inputs (Roberts et al., 2017; Mayer and Yang, 2022; Yang et al., 2024). The conversion models typically use several meteorological variables such as GHI, temperature, and wind speed as inputs, and require several calculation steps, hence the term "model chain", with individual models for the solar position, the separation of beam and diffuse irradiance, the shading loss, the PV performance, and other aspects (Yang et al., 2019; Wang et al., 2022). A variety of possible conversion models or components for individual processes exist and can be utilized to quantify forecast uncertainty by generating an ensemble of model chains (Mayer and Gróf, 2021; Mayer and Yang, 2022; Yang et al., 2024). Further, NWP models are typically run in ensemble mode by generating multiple simulation runs from varying initial conditions and/or changes to the model physics. This process yields a probabilistic forecast in the form of an ensemble, the members of which can be used as inputs to a model chain to generate an ensemble prediction of PV power (Wang et al., 2022). Ensembles of model chains can also be combined with ensemble forecasts from an NWP model to improve the uncertainty estimate (Mayer and Yang, 2023).

In the meteorological literature there is broad evidence that NWP ensemble predictions of various weather variables show systematic errors, which require correction to obtain accurate and reliable probabilistic forecasts. This correction process is called post-processing, for which an overview of common methods and recent developments can be found in Vannitsem et al. (2021). Most post-processing methods are statistical or machine learning (ML)-based distributional regression models where calibrated probabilistic forecasts are obtained in the form of parametric probability distributions, quantiles, or corrected ensemble predictions. One of the most popular post-processing methods is the ensemble model output statistics (EMOS; Gneiting et al., 2005) approach, where the forecast takes the form of a parametric distribution, the parameters of which are modelled as functions of summary statistics of the ensemble predictions. A recent focus of the post-processing literature has been the use of modern ML methods such as random forests (Taillardat et al., 2016) or neural networks (NNs; Rasp and Lerch, 2018), which allow for incorporating additional meteorological variables beyond the variable of interest as inputs, and have shown substantial improvements in predictive performance over classical statistical approaches such as EMOS [see, for example, Vannitsem et al. (2021) and Haupt et al. (2021) for overviews, and Demaeyer et al. (2023) for a benchmarking framework].

Statistical and ML-based post-processing methods have also been developed for the purpose of solar energy forecasting, most prominently for post-processing solar irradiance predictions from NWP models (e.g., Bakker et al., 2019; Le Gal La Salle et al., 2020; Yang and Gueymard, 2020; Yagli et al., 2020; Schulz et al., 2021; Yang and van der Meer, 2021; Baran and Baran, 2024; Song et al., 2024). Since similar post-processing methods can in principle be applied to the PV power predictions obtained as an output of the model chain, this allows for various ways of employing post-processing within probabilistic GHI-to-power conversion approaches utilizing model chains (Wang et al., 2022). As noted in related work on wind energy by Phipps et al. (2022), four different strategies are possible: not applying any post-processing at all and using the raw, unprocessed ensemble predictions obtained as outputs of the model chain (which we will denote by GHI^{raw}-PV^{raw}); applying post-processing only to the GHI predictions before the conversion (GHI^{pp}-PV^{raw}); applying post-processing only to the PV power forecasts obtained from the model chain conversion (GHI^{raw}-PV^{pp}); or applying post-processing in both steps (GHI^{pp}-PV^{pp}). Figure 1 provides a schematic overview of the different strategies. Using data-driven conversion models and statistical post-processing methods, Phipps et al. (2022) found in their study on wind energy that applying post-processing to the power forecasts is crucial to obtain accurate and reliable probabilistic forecasts.

The contributions of our work are threefold. First, we systematically evaluate the different strategies to assess the prospects of applying post-processing in model chain

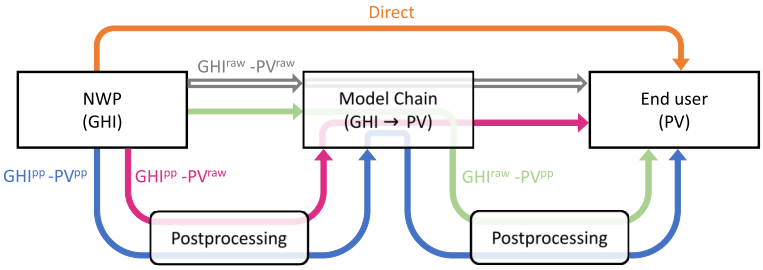

Fig. 1. Schematic illustration of the different strategies for applying post-processing methods within a model chain approach for PV power prediction.

approaches with a focus on solar energy, thus extending the work of Phipps et al. (2022). Theocharides et al. (2020) studied a similar question for deterministic forecasts, and we thus extend their work to probabilistic forecasts. Second, we propose NN-based post-processing methods for GHI and PV power forecasts, and systematically compare their performance to EMOS methods. Third, we compare the different strategies to an NN-based direct probabilistic PV power forecasting model, which uses the meteorological variables as inputs and produces probabilistic forecasts of PV power as its output without applying a model chain for the intermediate conversion step. Our study is based on a benchmark dataset for solar power forecasting (Wang et al., 2022) that comprises weather forecast and PV power observation data for a solar plant in the U.S.

The remainder of the article is organized as follows. Section 2 describes the benchmark dataset and additional data collection and pre-processing steps. Section 3 introduces the methods used for GHI and PV power post-processing and the models for the conversion from GHI to PV power. Results for the case study are presented in section 4, followed by a concluding discussion in section 5. Python code with implementations of all models to reproduce the results is available at https://github.com/HoratN/pp-modelchain.

2. Data

Our study is based on hourly data for the Jacumba Solar Project in southern California, U.S., covering the years 2017 to 2020. It comprises four different components, which will be introduced below. Three of these components are taken from Wang et al. (2022)[a]. For developing post-processing models, we use the data from 30 July 2017 to the end of 2019 as training data and the year 2020 as test data.

2.1. *Weather predictions*

The weather predictions include ensemble forecasts of GHI from the European Centre for Medium-Range Weather Forecasts (ECMWF), as well as deterministic predictions of additional weather variables from ECMWF's high-resolution (HRES) model. The GHI ensemble forecasts have 50 members and are initialized daily at 0000 UTC with a lead time of 24 hours. They contain hourly GHI averages in W m^{-2} and are time-stamped at the full hour marking the end of the averaging period. For the conversion from GHI to PV power we use a model chain approach that also takes temperature (°C) and wind speed (m s^{-1}) as inputs. Those variables will also be used as additional inputs to the NN-based post-processing models proposed in section 3. These forecasts contain instantaneous values at the end of the hour and therefore do not perfectly align with the remaining datasets. Note that HRES predictions of additional weather variables are available in the original dataset; however, here, we follow Wang et al. (2022) and restrict our attention to variables that are directly used within the model chain.

2.2. *GHI observations*

For the purpose of GHI post-processing, we require additional verifying data that can be used as ground truth observations. This observational dataset thus is the only part of the data used in our study that is not based on Wang et al. (2022). For this purpose, we use satellite-based GHI estimates, which we downloaded from the website of the National Solar Radiation Database (NSRDB; Sengupta et al., 2018)[b] and which contain hourly irradiance values in W m^{-2} for the location of the Jacumba solar plant (36.62°N, 116.13°W). We presume that the time stamps, e.g., [2017-01-01 00:30 UTC], correspond to the middle of the hourly averaging windows. To ensure consistency with the other datasets (i.e., the weather predictions and the PV power output observations), we therefore adjust the time stamps to the end of the averaging windows, resulting in [2017-01-01 01:00 UTC] for the previous example. Note that for the remainder of the article we refer to the satellite-based GHI estimates (and also the simulated PV power output introduced below) as "observational" data since they will be used as "best estimates" of the truth.

2.3. *PV power output observations*

Simulated hourly PV power output of the Jacumba solar plant in MW was published by the Lawrence Berkeley National Laboratory (Lawrence Berkeley National Lab. (2021); plant ID: 60947). The dataset contains PV power estimates computed with the System Advisor Model (SAM) by the National Renewable Energy Laboratory. The first available PV power data are recorded for 0000 UTC 30 July 2017 when the Jacumba solar project was put into operation. Since the data contain hourly PV values that are stamped at the beginning of the hour[c], we change the time stamp to the end of the averaging window to be consistent with the remaining datasets.

Note that these adjustments of the time stamps deviate from Wang et al. (2022); however, we found them to be helpful to better align the GHI and PV power observations. To illustrate this, Fig. 2 shows an exemplary day from all parts of the data for 21 January 2020. Due to the time stamp adjustments, the PV and GHI observations are well aligned in time and also match the diurnal cycle present in the ensemble GHI forecasts and the clear-sky GHI estimates. Note that here and in the remainder of the article, the hour of the day will always refer to the local time at the Jacumba solar plant in UTC-7h time.

3. Methods

Here, we briefly introduce relevant forecast evaluation methods used in our study, followed by descriptions of the model chain, and the different methods we propose for post-

[a] The data are available at https://github.com/wentingwang94/probabilistic-solar-forecasting
[b] https://nsrdb.nrel.gov/
[c] As detailed in the user guide available at https://live-etabiblio.pantheonsite.io/sites/default/files/user_guide_for_data_file.pdf

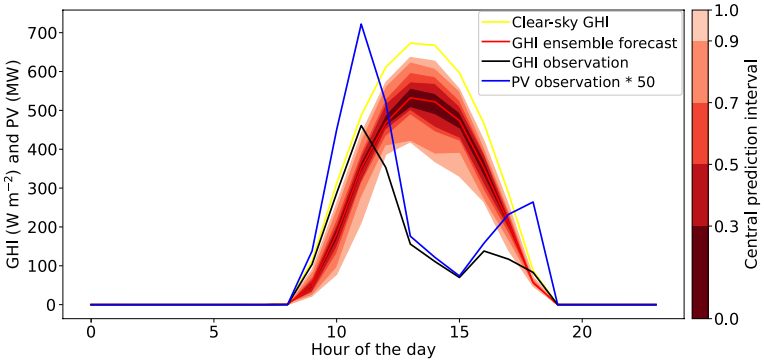

Fig. 2. Visualization of exemplary GHI ensemble predictions, GHI observations, and PV power observations at the Jacumba solar plant on 21 January 2020. Note that the PV power estimates are scaled by a factor of 50 to allow for a straightforward visual comparison with the GHI datasets, and that the clear-sky GHI values are included here for visualization purposes only.

processing the GHI and PV power forecasts, as well as the direct forecasting approach. Note that all implementation details are also available in the code provided in the replication code at https://github.com/HoratN/pp-modelchain. Results for the different strategies to apply post-processing within the model chain approach will be presented in section 4 below.

3.1. *Forecast evaluation*

Here, we provide a brief overview of the evaluation metrics employed, and refer to Lauret et al. (2019) for a detailed overview specifically tailored to probabilistic solar forecasts, as well as Gneiting et al. (2023a, section 4). It has now been widely accepted that probabilistic forecasts should be as sharp as possible, subject to being calibrated (Gneiting and Katzfuss, 2014). Calibration refers to the statistical consistency between the forecast distribution and the observation and essentially indicates whether the observation behaves like a random draw from the forecast distribution. To assess calibration, we use the histograms of the probability integral transform (PIT) $F(y)$, where F denotes the cumulative distribution function (CDF) of a probabilistic forecast, and y denotes the realizing observation. If the probabilistic forecast is calibrated, the PIT histogram should follow a uniform distribution and systematic deviations from uniformity can be used to identify misspecifications of the forecast distributions [see Gneiting et al. (2007) for details]. Note that censored forecast distributions with point masses in one or multiple points require adaptations to the calculation of the PIT value to account for the jumps in the forecast CDF. Here, we utilize the randomized PITs proposed in Czado et al. (2009). For probabilistic forecasts given in the form of an ensemble, verification rank histograms provide an analogous tool for visual calibration assessment. Thereby, the rank of the realizing observation when pooled with the ensemble predictions should be uniformly distributed. We further calculate the coverage and width of central prediction intervals (PIs) with nominal level $(m-1)/(m+1)$, where m is the size of the raw ensemble, which is 50 in our case. For a calibrated forecast, the coverage should be close to the nominal value, and the shorter the prediction interval, the sharper the forecast.

Further, proper scoring rules (Gneiting and Raftery, 2007) enable a simultaneous assessment of calibration and sharpness. The most widely used proper scoring rule in the meteorological literature is the continuous ranked probability score (CRPS; Matheson and Winkler, 1976),

$$\mathrm{CRPS}(F,y) = \int_{-\infty}^{\infty} (F(z) - \mathbb{I}\{y \leqslant z\})^2 \,\mathrm{d}z\,, \tag{1}$$

where F is a forecast CDF with a finite first moment, $\mathbb{I}$ is the indicator function, and y is the observation. Closed-form analytical expressions of the integral in Eq. (1) are available for a variety of parametric distributions, including the censored normal distributions we use below [see Jordan et al. (2019) for details].

To evaluate the accuracy of deterministic forecasts derived from the predictive distributions, we further consider the mean absolute error (MAE) given by

$$\mathrm{MAE}(F,y) = \frac{1}{n}\sum_{i=1}^{n} |x_i^{\mathrm{med}} - y_i|\,, \tag{2}$$

where x_i^{med} denotes the median of the forecast distribution F and y_i the observation, and n is the number of test samples, and the mean bias

$$\mathrm{Bias}(F,y) = \frac{1}{n}\sum_{i=1}^{n} \bar{x}_i - y_i\,, \tag{3}$$

where $\bar{x}_i$ is the mean value of the forecast distribution F.

We utilize Diebold–Mariano (DM) tests of equal predictive performance (Diebold and Mariano, 1995) to assess the statistical significance of score differences. Thereby, we conduct pairwise tests for each combination of two models, separately for all lead times. The test statistic of the DM test is given by

$$t = \sqrt{n}\frac{\bar{S}_n^F - \bar{S}_n^G}{\hat{\sigma}_n}, \tag{4}$$

where $\hat{\sigma}_n = \frac{1}{n}\sum_{i=1}^{n}[S(F_i,y_i) - S(G_i,y_i)]^2$.

Thereby, S denotes the scoring function, and $\bar{S}^F$ and $\bar{S}^G$ denote the corresponding mean scores for a fixed lead time for the two models' forecast distributions F and G and a corresponding test dataset of size n. Under the null hypothesis of equal predictive performance, the distribution of t approximately follows a standard Gaussian distribution. Note that here we assume independent forecast errors following Gneiting and Katzfuss (2014). In section 4 we present results of the DM test for the daylight hours from 0600 to 2000 local time (local time=UTC-7 hours) for a significance level of $\alpha = 0.05$.

3.2. *Model chain*

We employ a model chain approach to obtain ensemble forecasts of PV power and apply the model chain separately to each ensemble member. Since our main aim is not to find the best possible model chain configuration, but to study the role of post-processing in this context, we directly take the model chain setup from Wang et al. (2022) and implement it using code provided by the authors. It combines different component models (Erbs et al., 1982; Reindl et al., 1990; King et al., 2004; Reda and Andreas, 2004) to build the model chain [see Wang et al. (2022) for details]. Similar to Wang et al. (2022), we move the time stamp to the middle of the averaging period for applying the model chain, since the model chain also makes use of the time information for the estimation of the PV power output.

3.3. *Ensemble model output statistics*

As noted in the introduction, the EMOS approach proposed by Gneiting et al. (2005) is one of the most widely used post-processing methods in research and operations, and will serve as a baseline method for our comparisons. Phipps et al. (2022) applied EMOS to wind speed and wind power forecasts in a similar setting, albeit using data-driven conversion models. The EMOS approach relies on modelling the conditional distribution of the target variable Y, i.e., GHI or PV power in our case, given an ensemble of predictions $x_1,...,x_m$ of the target variable via a parametric probability distribution F_θ with parameters $\theta \in \mathbb{R}^d$, i.e.,

$$Y|x_1,...,x_m \sim F_\theta\,, \tag{5}$$

where $\theta = g(x_1,...,x_m)$ with a link function g that connects the distribution parameters with the ensemble prediction, typically via summary statistics such as the ensemble mean $\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$ or the ensemble variance $\mathrm{var}(x) = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2$. The choice of the parametric distribution F_θ plays a pivotal role in implementing EMOS models, and numerous extensions of the original normal-distribution-based EMOS model of Gneiting et al. (2005) from temperature and surface pressure to other meteorological variables have been proposed (e.g., in Thorarinsdottir and Gneiting, 2010; Lerch and Thorarinsdottir, 2013; Messner et al., 2014; Scheuerer, 2014; Baran and Lerch, 2015, 2016).

The EMOS models we apply here are based on censored normal distributions. For modelling GHI, several studies (e.g., Yang, 2020; Yagli et al., 2020; Le Gal La Salle et al., 2020) have used forecast distributions truncated at zero, where the probability mass of the negative values is redistributed to the positive values. Here, we follow Schulz et al. (2021) who proposed to instead use censored forecast distributions for GHI, where the probability mass of the negative values is added to zero as a point mass. This has the advantage that the same distribution can be used for all hours of the day, even during the night, when the GHI ensemble forecasts only contain zeros (Schulz et al., 2021). In contrast to Schulz et al. (2021) who applied a censored logistic distribution for GHI post-processing, we use a censored normal distribution. For GHI, we use a normal distribution that is left-censored at zero, and for PV power we use a doubly censored normal distribution with point masses at zero and 20 to restrict the output to the possible PV power production range in the dataset. The EMOS models for GHI and PV power both link the censored normal distributions (i.e., the location parameter μ and the scale parameter σ) to the corresponding ensemble forecasts of the target variable via

$$\begin{cases}\mu(x_1,...,x_m) = a + b\,\bar{x};\\ \sigma^2(x_1,...,x_m) = c + d\,\mathrm{var}(x)\,.\end{cases} \tag{6}$$

The EMOS parameters $a,b \in \mathbb{R}$ and $c,d \in \mathbb{R}_{\geq 0}$ are estimated by minimizing the mean CRPS over a training dataset. For formal definitions of the censored normal distribution and analytical formulas for computing the CRPS in closed form, see Jordan et al. (2019).

We consider two EMOS variants for both GHI and PV power. As a simple baseline, we train an EMOS model on data from all hours of the day by pooling all available training data into a single training dataset. This EMOS model thus applies the same correction to forecasts for every hour of the day and is not able to correct for hour-dependent errors, such as an underestimation of the PV production in the morning and an overestimation in the evening. To account for such diurnal variations, we further train separate EMOS models for every hour of the day, and refer to this approach as "EMOS hourly". By estimating separate sets of EMOS parameters for all hours of the day, these models thus have the advantage of being able to correct daytime-specific structures in the errors of the ensemble predictions, including systematic differences between day and night. A potential disadvantage of the EMOS hourly approach is that less training data are available for training the individual models. Lerch and Baran (2017) proposed alternative similarity-based estimation procedures for EMOS models that might be an interesting alternative for future studies.

3.4. *Neural network methods for post-processing*

Rasp and Lerch (2018) first proposed the use of NNs for probabilistic post-processing. The NN approach extends the EMOS framework by replacing the link function g with an NN, which connects the input predictors (e.g., summary statistics from the NWP ensemble) and the distribution parameters θ, which are obtained as the output of the NN. The

main advantages are that the NN enables the use of arbitrary input predictors, including ensemble predictions of other meteorological variables and exogeneous information, as well as the ability to flexibly model nonlinear dependencies between the inputs and the distribution parameters, which are learned in a data-driven way. Rasp and Lerch (2018) proposed a fully connected, feed-forward NN architecture, the parameters of which are optimized using the CRPS as a loss function. NN models for post-processing have been found to provide state-of-the-art predictive performance in many applications, and have been extended in various directions, including the use of alternative representations of the forecast distributions obtained as output of the NN (Bremnes, 2020; Schulz and Lerch, 2022; Song et al., 2024), or the use of more advanced NN architectures such as convolutional NNs that enable the incorporation of spatial information (Scheuerer et al., 2020; Veldkamp et al., 2021; Chapman et al., 2022; Horat and Lerch, 2024).

We consider two different variants of NN models for post-processing, which both use the ensemble prediction of GHI (i.e., the mean and standard deviation of the ECMWF ensemble) and deterministic forecasts of 2-m temperature and wind speed as inputs, but differ in the way they account for diurnal variations and treat the hour of the day. Analogous to the EMOS hourly model, we consider an NN model variant, where we train separate NN models for each hour of the day by using the corresponding subset of the training data only. This approach will be referred to as "NN hourly". As an alternative that more efficiently uses all available data, we further consider an NN model that is trained based on all available data comprising all hours of the day. To account for diurnal effects and differences over the different hours of the day, we provide the hour of the day information to the NN via embeddings, following a similar approach used by Rasp and Lerch (2018) to incorporate information about weather stations. Embeddings were originally proposed in natural language processing (Mikolov et al., 2013) and map categorical information to higher-dimensional latent representations in the form of vectors in $\mathbb{R}^p$. Rasp and Lerch (2018) used embeddings to incorporate information about the identifier of a weather station into an NN model for post-processing, which was then jointly trained over data from all locations, but made locally adaptive by using the latent representation obtained via the embeddings as additional inputs. Here, we follow a similar strategy and learn an embedding of the hour of the day to enable the NN model to learn how to exploit diurnal patterns in the input predictors[d]. We will refer to this model as the "NN" post-processing method[e]. In both NN approaches, we use a normal distribution that is left-censored at zero for GHI and a doubly censored normal distribution with point masses in zero and 20 for PV power as in the EMOS models, and estimate the weights of the NN by optimizing the CRPS as a loss function.

The NN architectures consist of two dense layers with 256 nodes each with ReLU activation functions, followed by one output layer with two nodes for the two distribution parameters. For the location parameter we use a linear activation, and employ a softplus activation function for the scale parameter to ensure positivity. For the NN with embeddings of the hour of the day, we concatenate the input forecasts with the output of the embedding layer that maps the hour of the day to a two-dimensional vector. Further, for GHI post-processing we replace the softplus activation function by a ReLU activation and add a small constant, i.e., $\mathrm{ReLU}(x) + 10^{-3}$, to improve numerical stability during training in light of point masses at zero during the night and occasional large outliers in the deviations between predicted and observed GHI during the day. All NN models are trained using the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 0.01 and early stopping, for which we use 20% of the training data as validation data, and restore the best weights from the previous epochs. For the hourly models we use a batch size of 256 and a patience of 5 epochs during the night hours (2300–0500 local time) and 30 epochs for the remaining hours to reduce the sensitivity to outliers during the day. On average, the models are trained for around 50 to 150 epochs, which is below the predefined maximum number of epochs. For the NN models with embeddings, we use a patience of 10 and a batch size of 1000, and train them for 50 epochs (however, the early stopping always terminates the training before reaching this limit). For both NN approaches we standardize the predictors. It is important to note that for the hourly NN model, the standardization is done separately for each hour and exclusively based on data from the specific hour. To account for the stochasticity of the training process, we repeat the model training 10 times for all NN models, and use the average of the distribution parameters from the 10 runs to obtain the final predictions.

3.5. *Direct forecasting model*

As an alternative to the model chain approach with post-processing, we further consider a direct forecasting method, where we use an NN to predict the PV power output directly from the available weather inputs, without the conversion via the model chain. For this purpose, we utilize the same NN model architectures as for post-processing the PV power forecasts. One particular advantage of the direct forecasting models is that they do not require any intricate domain knowledge or information about the specifications of the solar plant of interest. However, they require a training dataset of past weather predictions and corresponding PV power production to enable the model development, which is not the case for model chain approaches, at least if no post-processing is applied.

[d]Note that the use of embeddings ignores the temporal ordering of the hours of the day. A potential alternative for future research could be model architectures that directly use the difference of the hour of the day from the hour, where the maximum GHI value can be expected to be observed, as an input.

[e]However, note that in contrast to the baseline EMOS model, despite the lack of an "hourly" in the model name, this approach utilizes the hour of the day information.

4. Results

The results for the case study are presented in two parts. Section 4.1 focuses on the results of post-processing the GHI forecasts using EMOS or NN approaches. Then, section 4.2 evaluates the different strategies for employing post-processing methods in the model chain approach, and provides a comparison to the direct forecasting approach.

4.1. *GHI post-processing*

Figure 3 shows exemplary ECMWF ensemble forecasts of GHI along with corresponding observations. The observations often lie outside of the ensemble range, hence indicating that the ensemble spread is too small in the raw forecasts. Post-processing methods can increase the spread and thereby aim to improve calibration, as exemplified by the EMOS hourly forecasts illustrated in Fig. 3 alongside the ECMWF ensemble. While there is considerable day-to-day variability in the forecast uncertainty, the post-processed prediction bands typically entirely encompass those of the raw forecasts.

Table 1 presents evaluation scores for the ECMWF ensemble forecasts and all considered post-processing approaches. Results for pairwise tests of equal predictive performance to assess the statistical significance of the observed CRPS differences are provided in Table 2. As expected, post-processing shows substantial improvements over the raw ensemble forecasts; for example, of up to 25% in terms of the mean CRPS and around 18% in terms of the MAE. Clear improvements are also achieved in terms of coverage of central prediction intervals. While the raw ECMWF ensemble predictions provide the sharpest forecasts with the shortest prediction intervals, they lack calibration since on average, only 31.1% of the daylight observations lie within the 96.1% prediction interval. By contrast, all post-processing methods provide better calibrated forecasts and achieve coverages close to the nominal level, with substantially wider prediction intervals.

All post-processing methods perform more or less equally well, with the hourly approaches achieving slightly better scores than the methods trained on data from the entire day. With the exception of the EMOS approach, all methods significantly outperform the raw ECMWF predictions at most daylight hours. The NN approaches outperform the EMOS variants slightly, even though the relative differences between NN hourly and EMOS hourly are less than 1% in terms of the mean CRPS. Similar conclusions apply to the other evaluation metrics. The observed CRPS differences between the NN approaches and EMOS hourly tend to be significant in favor of the NN variants for twice the daylight hours than vice versa (see Table 2). The differences

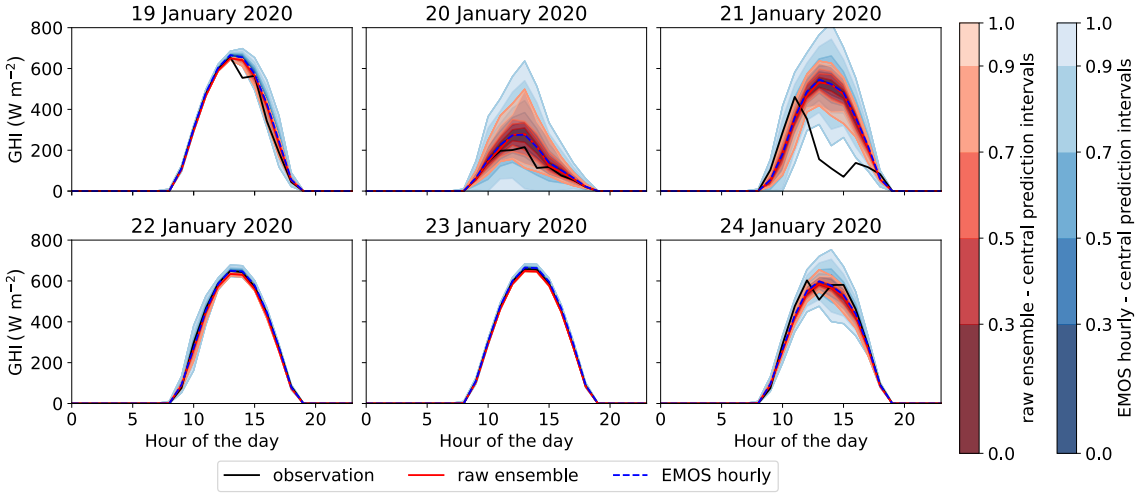

Fig. 3. Exemplary probabilistic GHI forecasts based on the ECMWF ensemble and the EMOS hourly post-processing approach for dates in January 2020. The coloured areas indicate central prediction intervals. The lines show the ensemble median.

Table 1. Mean values of various evaluation metrics for GHI forecasts from the raw ECMWF ensemble and all considered post-processing methods, averaged over all 24 hours of the day and all days in the test dataset. For the PI coverage and the PI width, averages across daylight hours, i.e., from 6:00–20:00 local time, are also shown. PI coverage and width are computed for central prediction intervals with the nominal coverage of the raw ensemble, i.e., $(m-1)/(m+1)$ for $m = 50$, which is approximately 96.1%. The best methods are indicated in bold.

	CRPS	MAE	PI Cover.	PI Cover. daytime	PI Width	PI Width daytime
ECMWF	14.676	17.337	57.0	31.1	28.0	44.7
EMOS	11.510	14.851	91.3	86.2	68.2	105.7
EMOS hourly	11.080	**14.282**	**95.4**	**92.5**	65.8	105.2
NN	11.262	14.725	94.0	90.4	66.4	106.3
NN hourly	**11.046**	14.473	94.8	91.6	64.4	103.0

between the two NN models are not significant in terms of CRPS for most daylight hours.

In contrast to other applications of NN methods for post-processing, the improvements achieved by the NN models over the EMOS approaches are on a notably smaller scale. However, they are in line with previous findings indicating that the main advantage of using NN methods for post-processing is the efficient use of additional input information (Rasp and Lerch, 2018). The additional inputs available to the NN models here likely do not provide substantial predictive information about GHI, and the main advantage of the NN approaches over the EMOS models thus might be given by the potential to learn nonlinear link functions [see also Demaeyer et al. (2023) for related results].

Figure 4 shows the evolution of various evaluation metrics over the course of the day. As expected, all evaluation metrics show a strong dependence on the time of the day, and the CRPS, bias and width of the prediction intervals are almost zero for all methods during night time. During night time, all forecasts coincide at zero, which prevents a proper computation of coverage and PI width, since the coverage equals one for all levels, and the corresponding PI width is always zero. The CRPS curves of all forecasts show a maximum at around hour 14, and the curves for almost all post-processing methods lie below the CRPS curve of the raw ensemble for almost all hours of the day. The only notable exception is the NN model at hour 20, which likely corresponds to numerical stability issues in the parameter estimation as this outlier is also present in the bias and coverage curves.[f] Interestingly, the CRPS curves for EMOS, i.e., the only post-pro-

Table 2. Statistical significance of the CRPS differences between the different post-processing methods for GHI and the corresponding ECMWF forecasts. The values indicate the number of daylight hours, from a total of 15, for which the model in the row performs significantly better than the model in the column in terms of CRPS. The significance is computed with pairwise two-sided DM tests at a significance level of $\alpha = 0.05$, as detailed in section 3.1.

	ECMWF	EMOS	EMOS hourly	NN	NN hourly
ECMWF	–	7	0	0	0
EMOS	8	–	0	1	1
EMOS hourly	14	12	–	4	4
NN	14	11	8	–	2
NN hourly	14	9	7	1	–

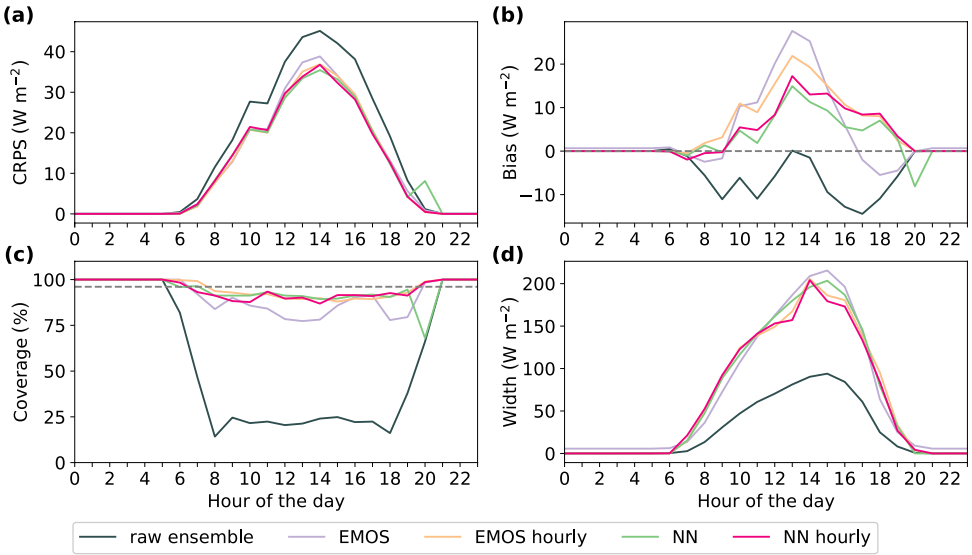

Fig. 4. Hourly values of the CRPS (a), and the bias (b) of the mean forecast, as well as the coverage (c) and width (d) of 96.1% prediction intervals for the GHI forecasts. All values are averaged over the test dataset.

[f] A more detailed investigation suggests that this might be due to notable violations from the distributional assumptions, indicated by heavily skewed histograms of the GHI observations at this hour with most observations at 0, but a long tail with values up to 50. While such distributions might be challenging to model with a censored normal distribution in general, the estimation of the NN models seems to show particular difficulties in converging to reasonable parameter estimates. Non-parametric methods such as Bernstein Quantile Networks (Bremnes, 2020; Gneiting et al., 2023a) or quantile regression (Song et al., 2024) could provide a possible remedy by also allowing for non-symmetric and non-normal distributions. For example, Gneiting et al. (2023a) showed examples comparing NN learning distributional parameters and BQN methods and reported better CRPS scores for the BQN method than for the parametric NN approaches for GHI due to the enhanced flexibility of the BQN method.

cessing model that does not use information about the hours of the day, closely follows the CRPS curves of the other post-processing models except for hours around midday, where the model shows a larger CRPS. The MAE curves look almost identical to those of the CRPS and thus are omitted here. All post-processing models show a slightly positive bias with a maximum around midday. Since the raw ensemble shows a negative bias throughout the day, the sign (but not the magnitude) of the bias changes due to post-processing. The NN approaches show smaller biases during most of the day, while the bias structure of the EMOS model reveals that the same bias correction is applied to all hours of the day and the model is unable to learn hour-dependent error characteristics. However, given the magnitude of the bias of all methods, the improvements in terms of the CRPS largely stem from an improved calibration of the forecasts. This also becomes evident from the coverage and width of the prediction intervals from the different methods. Not surprisingly, the raw ensemble forecasts produce the shortest prediction intervals and thus the sharpest forecasts throughout the day; however, they fail to achieve a coverage close to the nominal value. All post-processing methods yield substantially wider prediction intervals and a better coverage. Again, the EMOS model shows a slightly worse performance than all other methods that use predictive information about the hour of the day. In terms of the prediction interval width, it is interesting to note that the hourly EMOS and NN methods yield slightly sharper forecasts than their corresponding more general counterparts. A closer inspection of the performance of the EMOS model during night time indicates that neither the bias nor the prediction interval width is exactly zero. This approach thus fails to appropriately model the GHI values during night times as a point mass in zero.

To further assess the calibration of the probabilistic forecasts, Fig. 5 shows verification rank and PIT histograms of all approaches for selected hours of the day. The raw ECMWF ensemble forecasts of GHI are clearly underdispersive and lack calibration, as indicated by the U-shaped verification rank histograms. The PIT histograms of all post-processed forecasts are notably closer to the desired uniform distribution, and thus indicate that these forecasts are better calibrated. The EMOS hourly and the two NN approaches show the best calibration. By contrast, the EMOS model jointly estimated over all hours of the day produces less well calibrated forecasts and a clear bias in the form of an overestimation of the GHI values at hour 10, and an underestimation at hour 18, respectively.

4.2. *PV power forecasts*

Here, we first present the results for the different strategies of applying post-processing in a model chain approach, and then compare them to a direct forecasting model.

4.2.1. *Post-processing in the model chain approach*

Table 3 shows mean CRPS values for all combinations of strategies for applying post-processing and post-processing methods. All combinations improve the PV power forecasts compared to using the model chain approach without any

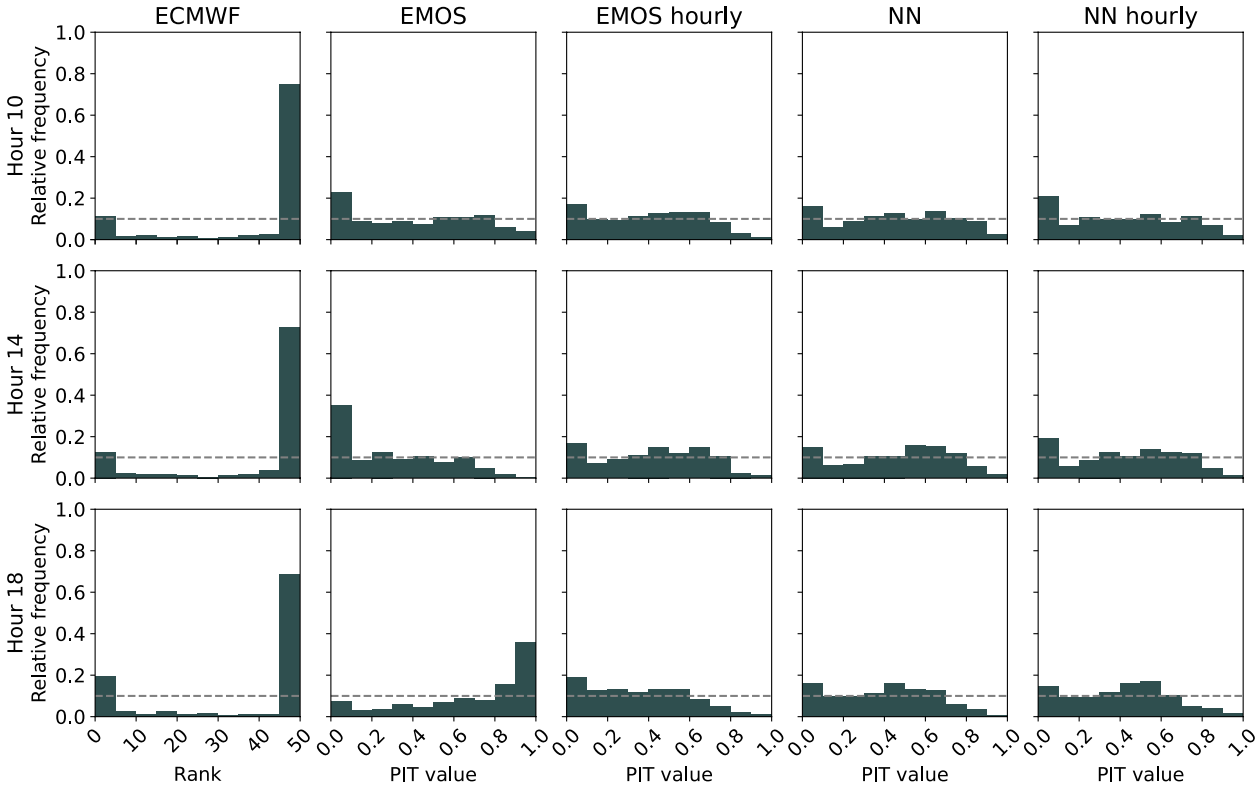

Fig. 5. Verification rank histogram of the ECMWF ensemble forecasts and PIT histograms for all considered post-processing methods for GHI for hours 10, 14, and 18.

Table 3. Mean CRPS values for probabilistic PV power forecasts obtained from the considered post-processing strategies using different post-processing methods. Note that in the GHI^{pp}-PV^{pp} strategy, the same post-processing method is applied in both steps. All CRPS values are averaged over all hours of the day. The model chain approach without any post-processing (GHI^{raw}-PV^{raw}) achieves a CRPS of around 0.689. The best methods are indicated in bold.

Strategy	EMOS	EMOS hourly	NN	NN hourly
GHI^{pp}-PV^{raw}	0.676	0.651	0.639	0.644
GHI^{raw}-PV^{pp}	0.564	0.305	**0.294**	0.309
GHI^{pp}-PV^{pp}	0.573	0.306	**0.294**	0.308

post-processing, but the magnitude of the improvements differs substantially across methods and strategies. Applying post-processing only to the GHI forecasts in the GHI^{pp}-PV^{raw} strategy leads to the smallest improvements of at most around 7% in terms of the mean CRPS. For all post-processing methods, applying post-processing to the PV power forecasts obtained from the model chain appears to be the most crucial step, as the GHI^{raw}-PV^{pp} and GHI^{pp}-PV^{pp} strategies achieve almost identical mean CRPS values for all post-processing methods, and improvements over the raw model chain forecasts of up to around 57%. The GHI^{pp}-PV^{pp} and GHI^{raw}-PV^{pp} strategies (excluding EMOS) are significantly better than the GHI^{pp}-PV^{raw} strategy for many of the daylight hours (not shown).

In terms of the different post-processing methods, the EMOS model jointly estimated for all hours of the day performs substantially worse than all others. The best overall CRPS values for all strategies are achieved by the NN model that uses the hour of the day information via embeddings. The improvements over the hourly models for GHI^{pp}-PV^{pp} and GHI^{raw}-PV^{pp} are significant for one third of all daylight hours. In contrast to the GHI forecasting task, where this model performed worse than the two hourly approaches, it thus might be more beneficial for PV power forecasting to utilize a NN with an increased training sample size. The two hourly models, EMOS hourly and NN hourly, achieve very similar CRPS values, with slightly better scores for the simpler EMOS hourly model. However, for most daylight hours the differences between EMOS hourly and NN hourly are not significant, neither for the GHI^{pp}-PV^{pp} nor the GHI^{raw}-PV^{pp} strategy. As discussed in the results for the GHI predictions, the benefits of using an NN approach here might again be limited by the information content of the additional predictors available to the NNs. Further, the EMOS model is notably simpler to tune, with fewer hyperparameters and optimization settings that need to be chosen.

To assess the diurnal variability of the forecast errors, Fig. 6 shows hourly values of the mean CRPS and bias for all combinations of strategies and post-processing methods. Note that the figure also contains results for the direct forecasts, which we will discuss below. Both the model chain approach based on the raw ECMWF ensemble predictions without any post-processing, as well as the GHI^{pp}-PV^{raw} strategy, independent of the post-processing method, show substantially larger forecast errors with a pronounced diurnal cycle. The largest forecast errors occur during the early morning and late afternoon hours, whereas the CRPS during midday is comparable to that of the other strategies and post-processing methods. A similar behaviour, albeit with smaller errors in the morning, but larger errors in the afternoon, can be observed for the EMOS approach estimated jointly over all hours of the day in the two remaining strategies (GHI^{raw}-PV^{pp} and GHI^{pp}-PV^{pp}). The main explanation for these observations is likely the behaviour of the bias, which is notably larger than it was for the GHI forecasts when compared to the magnitude of the CRPS. The less well performing strategies and methods (i.e., the raw model chain without post-processing, the GHI^{pp}-PV^{raw} strategy, and all non-hourly EMOS variants) show substantially larger biases that change the sign from positive (i.e., overestimation) in the morning to negative (i.e., underestimation) in the afternoon. Not surprisingly, the EMOS model that is jointly estimated over all hours of the day is not able to account for these daytime-specific variations in the bias. All other combination, where post-processing is applied to the PV power forecasts obtained as the output of the conversion model and information about the hour of the day enters the model, achieve notably smaller biases and lower CRPS values during the whole day. The relative differences between these approaches are only minor, with slightly increased biases of the EMOS hourly forecasts during the afternoon.

Due to the large number of combinations of strategies and methods, the following graphical illustrations and corresponding discussions focus on results for the best-performing strategy (GHI^{pp}-PV^{pp}) and/or method (NN). Figure 7 shows exemplary probabilistic PV power forecasts and corresponding observations for the model chain without any post-processing and the GHI^{pp}-PV^{pp} strategy with the NN model for post-processing for the same days as the GHI forecasts in Fig. 3. A large variability in the forecast uncertainty can be observed over the different days, which seems to be directly connected to the forecast uncertainty of the corresponding GHI predictions. As for GHI, post-processing here substantially increases the width of the prediction intervals.

Figure 8 shows verification rank and PIT histograms for the GHI^{raw}-PV^{raw} and GHI^{pp}-PV^{pp} strategies and all considered post-processing methods. In light of the biases observed in Fig. 6, it is not surprising that neither the model chain without any post-processing nor the GHI^{pp}-PV^{pp} strategy with EMOS post-processing show calibrated forecasts, but clearly notable biases, in particular during hours

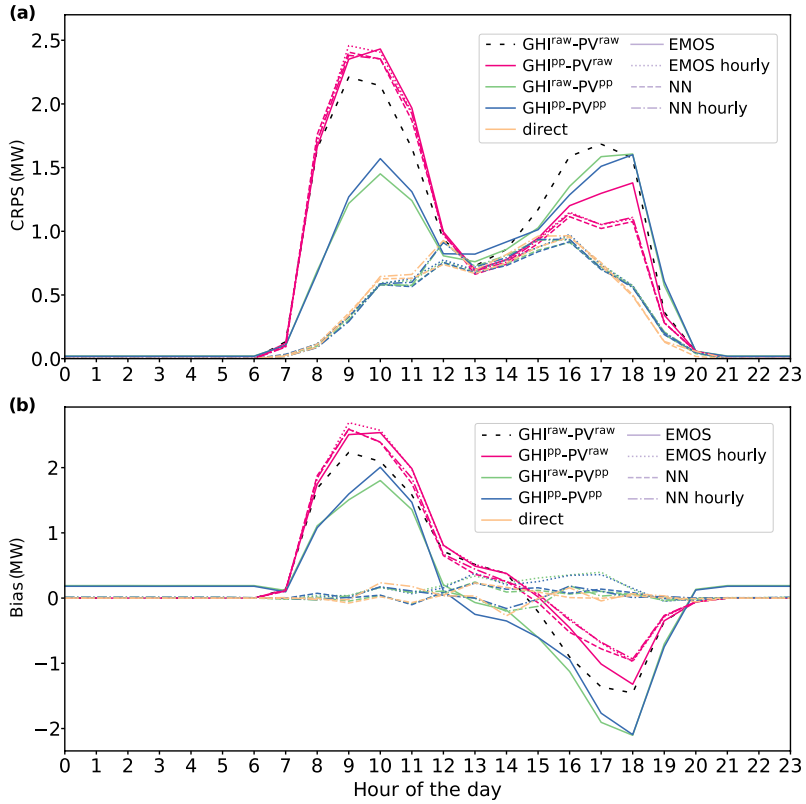

Fig. 6. Hourly values of the mean CRPS (a) and the bias (b) of the mean forecast for the considered strategies and post-processing methods, as well as the direct forecasts. Note that the strategies are indicated by line type, and the post-processing methods by the colour of the corresponding lines.

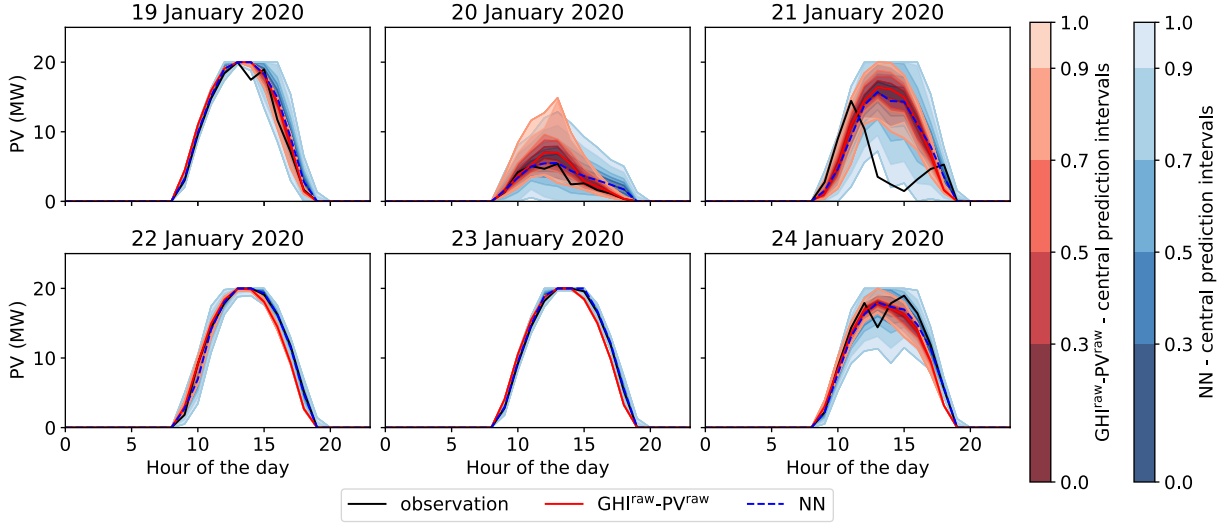

Fig. 7. Exemplary probabilistic PV forecasts for dates in January 2020 for GHI^{raw}-PV^{raw} and GHI^{pp}-PV^{pp} with the NN jointly estimated for all hours of the day. The coloured areas indicate central prediction intervals. The lines show the ensemble median.

10 and 18. All other post-processing methods yield substantially better calibrated forecasts with PIT histograms much closer to uniformity.

4.2.2. *Direct forecasting model*

Table 4 shows various evaluation metrics for the NN-based direct forecasting methods for PV power that do not utilize a conversion from GHI to PV power via the model chain. Results for corresponding pairwise DM tests of equal predictive performance in terms of the CRPS are provided in Table 5. Both the direct NN model jointly estimated for all hours via embeddings and the hourly direct NN model

show substantially better CRPS values than the model chain approach without any post-processing. Pairwise DM tests show that the NN models significantly improve the predictions for most of the daylight hours. The mean CRPS values achieved by the direct forecasting models are comparable to those of their GHI^{pp}-PV^{pp} counterparts, albeit slightly worse. Revisiting Fig. 6, we note that the diurnal pattern of the direct models is almost identical to that of the corresponding GHI^{pp}-PV^{pp} model. The two approaches are also very similar in terms of coverage and PI width. For the daylight

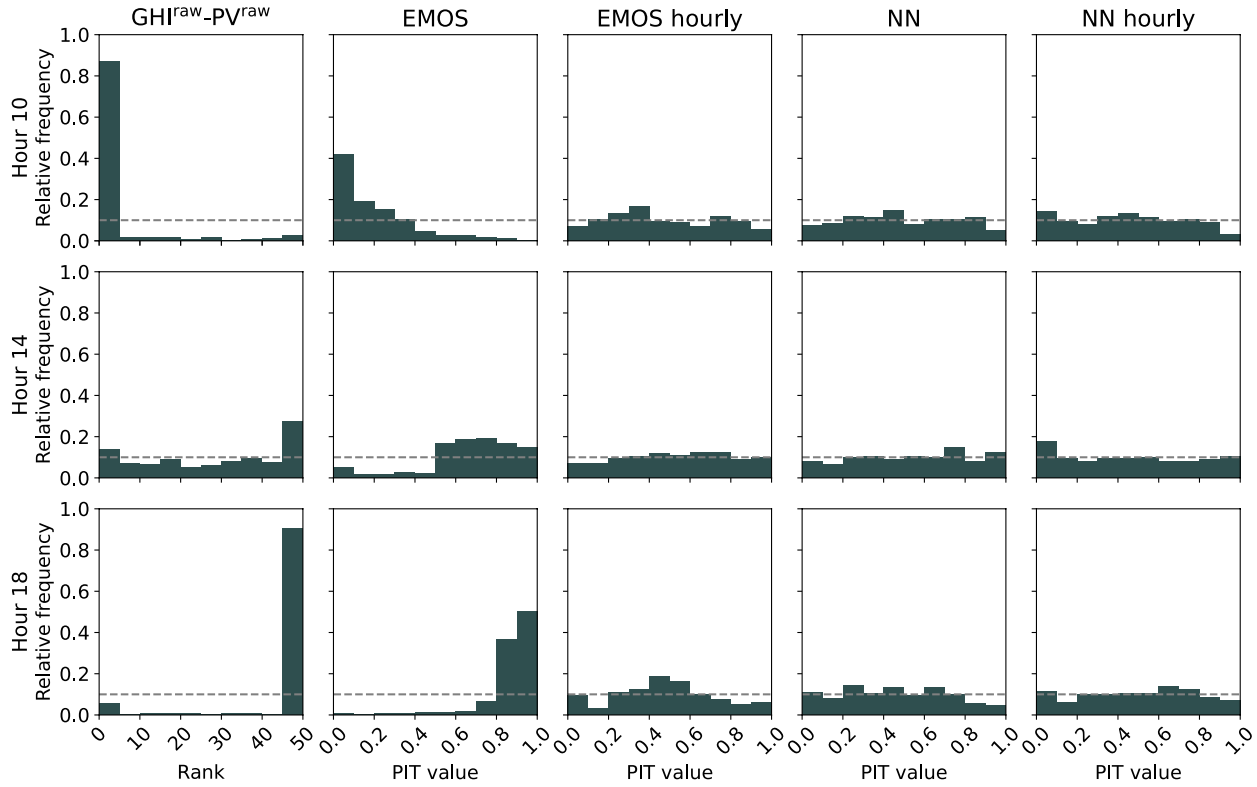

Fig. 8. Verification rank histogram and PIT histograms for the raw forecasts (GHI^{raw}-PV^{raw}) and all considered post-processing methods for PV power output within the GHI^{pp}-PV^{pp} approach for hours 10, 14, and 18.

Table 4. Mean CRPS values for probabilistic PV power forecasts obtained from the GHI^{pp}-PV^{pp} post-processing strategies using NN methods in comparison to NN-based direct forecasting methods that do not utilize a conversion from GHI to PV power via the model chain. Scores are averaged over all 24 hours of the day and all days in the test dataset. For the PI coverage and the PI width, averages across daylight hours, i.e., from 6:00–20:00 local time, are also shown. PI coverage and width are computed for PIs with the nominal coverage of the raw ensemble, i.e., $(m-1)/(m+1)$ for $m = 50$, which is approximately 96.1%. The best methods are indicated in bold.

	CRPS	MAE	PI Cover.	PI Cover. daytime	PI Width	PI Width daytime
GHI^{raw}-PV^{raw}	0.698	0.768	65.4	44.7	0.686	1.098
GHI^{pp}-PV^{pp} NN	**0.294**	**0.397**	97.5	**96.0**	1.808	2.894
GHI^{pp}-PV^{pp} NN hourly	0.308	0.413	97.8	96.4	2.773	4.320
Direct NN	0.298	0.409	97.9	96.7	1.841	2.944
Direct NN hourly	0.314	0.422	**96.1**	93.8	2.535	4.048

Table 5. Statistical significance of the CRPS differences for the methods listed in Table 4. The values indicate the number of daylight hours, from a total of 15, for which the model in the row performs significantly better than the model in the column in terms of CRPS. The significance is computed with pairwise two-sided DM tests at a significance level of $\alpha = 0.05$, as detailed in section 3.1.

	GHI^{raw}-PV^{raw}	GHI^{pp}-PV^{pp} NN	GHI^{pp}-PV^{pp} NN hourly	Direct NN	Direct NN hourly
GHI^{raw}-PV^{raw}	–	0	1	0	0
GHI^{pp}-PV^{pp} NN	13	–	4	6	6
GHI^{pp}-PV^{pp} NN hourly	11	3	–	5	4
Direct NN	13	4	5	–	4
Direct NN hourly	10	4	3	4	–

hours, all NN approaches achieve a coverage that is very close to the nominal coverage of 96.1%. However, the hourly models have prediction intervals that are up to 50% wider than the models estimated on data from all hours of the day. An investigation of the width of the prediction intervals for every hour of the day (not shown) reveals that the PIs for 13:00 and 14:00 local time are extremely wide, reaching up to 18 MW. As discussed above, we assume that for these hours, the assumption of a doubly censored normal distribution is violated since the distribution of the observed PV power is heavily skewed.

5. Discussion and conclusions

We have systematically compared different strategies for employing post-processing to improve probabilistic PV power forecasts within a model chain approach, where weather predictions are converted to PV power via a cascade of physics-based models. In a case study for a solar plant in the U.S. based on data from Wang et al. (2022), we develop statistical and ML methods for post-processing GHI and PV power forecasts. We find that post-processing leads to substantial improvements when applied to the PV power forecasts that are obtained as the output of the model chain, in line with findings from Phipps et al. (2022) in the context of wind energy prediction. Whether or not the GHI forecasts are post-processed before using them as input to the model chain plays an almost negligible role for most post-processing methods.

In terms of the performance of the different post-processing approaches, the use of the hour of the day is of central importance when building a model, either via utilizing separate models for each hour of the day, or by including the temporal information as input to an NN model, in our case via embeddings. Comparing classical EMOS and modern NN-based models for post-processing, we find that in contrast to various recent studies on other weather variables, the use of NNs here only leads to minor improvements. A likely explanation of this finding is that the additional input information that was available to the NN models carries too little predictive information to be effectively utilized, since we restricted our attention to those meteorological variables that also served as an input to the model chain (i.e., deterministic temperature and wind speed predictions).

We have further proposed an NN model for directly predicting PV power output from the weather information without using a model chain for the intermediate conversion of GHI to PV power. This direct forecasting model showed almost competitive forecast performance with the best combination of post-processing strategy and method in the model chain setting, but comes with the advantage of being generally applicable without requiring specific knowledge about the individual solar plant's design and technical specifications. However, estimating a direct forecasting model of course requires past weather predictions and PV power observations as training data, which is not necessary for a model chain approach, at least if no post-processing is applied. Regarding the practical implementation of the considered forecasting methods, both the direct as well as the post-processing approach likely constitute viable options and the best choice for a particular application will depend on the specific knowledge about technical specifications and the data availability. An interesting question for future research might be whether there are differences in terms of the requirements on the amount of available past data to achieve similar predictive performance in both approaches.

Our study provides several avenues for further model development and analysis. Perhaps most importantly, it only constitutes a first step towards a more systematic analysis and comparison of the different strategies and post-processing methods, since we only used data from a single solar plant and a single model chain approach. Repeating the analysis on a more comprehensive dataset with multiple locations and potentially an ensemble of model chains (Mayer and Yang, 2022) would not only provide a more comprehensive comparison, but would also allow for addressing interesting methodological questions, e.g., how to effectively develop NN model architectures for multiple sites, or how to post-process multi-model ensembles in this setting.

Another natural starting point for future research are ways to further improve our NN models for post-processing and direct PV power prediction. Instead of learning distribution parameters with the NNs, non-parametric methods would allow for non-symmetric and non-normal distributions of the target variable and could provide a better fit for the heavily skewed distributions of GHI and PV power during midday, early morning or late evening hours [see Bremnes (2020), Schulz and Lerch (2022), and Gneiting et al. (2023a) for examples]. Further, using deterministic predictions of more weather variables from the data available in Wang et al. (2022) as inputs to the NN models might lead to improvements. Note that we only post-processed the GHI predictions, but not the other inputs to the model chain and the NN models. If observations for those variables were available, post-processed forecasts might further improve the performance of the GHI to PV power conversion via the model chain. In all of these developments, it would furthermore be interesting to consider additional aspects of forecast quality beyond statistical evaluation metrics, such as economic aspects (Van der Meer et al., 2018; Gneiting et al., 2023b).

Finally, over the past few years, there have been rapid developments in AI-based data-driven weather models such as Pangu-Weather (Bi et al., 2023), GraphCast (Lam et al., 2023) or AIFS (Lang et al., 2024). Since those models have been demonstrated to outperform classical physics-based NWP models on a variety of prediction tasks, investigating whether they could also replace NWP models as inputs for solar energy prediction might be an interesting question for future research. The use of post-processing methods investigated in our article might play an important role in this context, since these AI weather models likely exhibit different systematic error characteristics, and model chains adapted to physics-based inputs thus might not work as well. Fur-

ther, many currently available models do not provide relevant outputs like GHI[h] and are limited to deterministic predictions only. They thus require additional steps for quantifying forecast uncertainties, which are particularly relevant in applications such as solar energy prediction (Bülte et al., 2024).

Acknowledgements. The research leading to these results was carried out within the Young Investigator Group "Artificial Intelligence for Probabilistic Weather Forecasting" funded by the Vector Stiftung. In addition, this project received funding from the Federal Ministry of Education and Research (BMBF) and the Baden-Württemberg Ministry of Science as part of the Excellence Strategy of the German Federal and State Governments. We thank Peter Knippertz, Wenting Wang and Dazhi Yang for helpful comments and discussions. We further thank the two anonymous reviewers, whose constructive comments helped to improve an earlier version of this paper.

Funding note Open Access funding enabled and organized by Projekt DEAL.

REFERENCES

Appino, R. R., J. Á. González Ordiano, R. Mikut, R. Faulwasser, and V. Hagenmeyer, 2018: On the use of probabilistic forecasts in scheduling of renewable energy sources coupled to storages. *Applied Energy*, **210**, 1207–1218, https://doi.org/10.1016/j.apenergy.2017.08.133.

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

Baran, Á., and S. Baran, 2024: A two-step machine-learning approach to statistical post-processing of weather forecasts for power generation. *Quart. J. Roy. Meteor. Soc.*, **150**, 1029–1047, https://doi.org/10.1002/qj.4635.

Baran, S., and S. Lerch, 2015: Log-normal distribution based ensemble model output statistics models for probabilistic wind-speed forecasting. *Quart. J. Roy. Meteor. Soc.*, **141**, 2289–2299, https://doi.org/10.1002/qj.2521.

Baran, S., and S. Lerch, 2016: Mixture EMOS model for calibrating ensemble forecasts of wind speed. *Environmetrics*, **27**, 116–130, https://doi.org/10.1002/env.2380.

Bi, K. F., L. X. Xie, H. H. Zhang, X. Chen, X. T. Gu, and Q. Tian, 2023: Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, **619**, 533–538, https://doi.org/10.1038/s41586-023-06185-3.

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

Bülte, C., N. Horat, J. Quinting, and S. Lerch, 2024: Uncertainty quantification for data-driven weather models. Available from https://arxiv.org/abs/2403.13458 .

Chapman, W. E., L. Delle Monache, S. Alessandrini, A. C. Subramanian, F. M. Ralph, S.-P. Xie, S. Lerch, and N. Hayatbini, 2022: Probabilistic predictions from deterministic atmospheric river forecasts with deep learning. *Mon. Wea. Rev.*, **150**, 215–234, https://doi.org/10.1175/MWR-D-21-0106.1.

Czado, C., T. Gneiting, and L. Held, 2009: Predictive model assessment for count data. *Biometrics*, **65**, 1254–1261, https://doi.org/10.1111/j.1541-0420.2009.01191.x.

Demaeyer, J., and Coauthors, 2023: The EUPPBench postprocessing benchmark dataset v1.0. *Earth System Science Data*, **15**, 2635–2653, https://doi.org/10.5194/essd-15-2635-2023.

Diebold, F. X., and R. S. Mariano, 1995: Comparing predictive accuracy. *Journal of Business & Economic Statistics*, **13**, 253–263, https://doi.org/10.1080/07350015.1995.10524599.

Erbs, D. G., S. A. Klein, and J. A. Duffie, 1982: Estimation of the diffuse radiation fraction for hourly, daily and monthly-average global radiation. *Solar Energy*, **28**, 293–302, https://doi.org/10.1016/0038-092X(82)90302-4.

Fraunhofer Institute for Solar Energy Systems, 2024: Recent facts about photovoltaics in Germany. Technical report. Available from https://www.ise.fraunhofer.de/en/publications/studies/recent-facts-about-pv-in-germany.html.

Gneiting, T., and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102**, 359–378, https://doi.org/10.1198/016214506000001437.

Gneiting, T., and M. Katzfuss, 2014: Probabilistic forecasting. *Annual Review of Statistics and its Application*, **1**, 125–151, https://doi.org/10.1146/annurev-statistics-062713-085831.

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

Gneiting, T., F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **69**, 243–268, https://doi.org/10.1111/j.1467-9868.2007.00587.x.

Gneiting, T., S. Lerch, and B. Schulz, 2023a: Probabilistic solar forecasting: Benchmarks, post-processing, verification. *Solar Energy*, **252**, 72–80, https://doi.org/10.1016/j.solener.2022.12.054.

Gneiting, T., and Coauthors, 2023b: Model diagnostics and forecast evaluation for quantiles. *Annual Review of Statistics and its Application*, **10**, 597–621, https://doi.org/10.1146/annurev-statistics-032921-020240.

Gottwalt, S., J. Gärttner, H. Schmeck, and C. Weinhardt, 2017: Modeling and valuation of residential demand flexibility for

[h] A notable exception is the recently proposed FuXi-2.0 model (Zhong et al., 2024), which explicitly aims at solar and wind energy forecasting.

renewable energy integration. *IEEE Transactions on Smart Grid*, **8**, 2565–2574, https://doi.org/10.1109/tsg.2016.2529424.

Haupt, S. E., and Coauthors, 2019: The use of probabilistic forecasts: Applying them in theory and practice. *IEEE Power and Energy Magazine*, **17**, 46–57, https://doi.org/10.1109/MPE.2019.2932639.

Haupt, S. E., W. Chapman, S. V. Adams, C. Kirkwood, J. S. Hosking, N. H. Robinson, S. Lerch, and A. C. Subramanian, 2021: Towards implementing artificial intelligence post-processing in weather and climate: Proposed actions from the Oxford 2019 workshop. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **379**, 20200091 https://doi.org/10.1098/rsta.2020.0091.

Horat, N., and S. Lerch, 2024: Deep learning for postprocessing global probabilistic forecasts on subseasonal time scales. *Mon. Wea. Rev.*, **152**, 667–687, https://doi.org/10.1175/MWR-D-23-0150.1.

Jordan, A., F. Krüger, and S. Lerch, 2019: Evaluating probabilistic forecasts with scoringRules. *Journal of Statistical Software*, **90**, 1–37, https://doi.org/10.18637/jss.v090.i12.

King, D. L., W. E. Boyson, and J. A. Kratochvil, 2004: Photovoltaic array performance model. AC04-94AL85000.

Kingma, D. P., and J. Ba, 2015: Adam: A method for stochastic optimization. *Proc. 3rd International Conf. on Learning Representations*, San Diego, USA.

Lam, R., and Coauthors, 2023: Learning skillful medium-range global weather forecasting. *Science*, **382**, 1416–1421, https://doi.org/10.1126/science.adi2336.

Lang, S., and Coauthors, 2024: AIFS – ECMWF's data-driven forecasting system. Available from https://arxiv.org/abs/2406.01465.

Lauret, P., M. David, and P. Pinson, 2019: Verification of solar irradiance probabilistic forecasts. *Solar Energy*, **194**, 254–271, https://doi.org/10.1016/j.solener.2019.10.041.

Lawrence Berkeley National Lab., 2021: Solar-to-grid public data file for utility-scale (UPV) and distributed photovoltaics (DPV) generation, capacity credit, and value for 2012-2020. Available from https://dx.doi.org/10.25984/1825661 .

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

Lerch, S., and T. L. Thorarinsdottir, 2013: Comparison of non-homogeneous regression models for probabilistic wind speed forecasting. *Tellus A: Dynamic Meteorology and Oceanography*, **65**, 21206, https://doi.org/10.3402/tellusa.v65i0.21206.

Lerch, S., and S. Baran, 2017: Similarity-based semilocal estimation of post-processing models. *Journal of the Royal Statistical Society Series C: Applied Statistics*, **66**, 29–51, https://doi.org/10.1111/rssc.12153.

Matheson, J. E., and R. L. Winkler, 1976: Scoring rules for continuous probability distributions. *Management Science*, **22**, 1087–1096, https://doi.org/10.1287/mnsc.22.10.1087.

Mayer, M. J., and G. Gróf, 2021: Extensive comparison of physical models for photovoltaic power forecasting. *Applied Energy*, **283**, 116239, https://doi.org/10.1016/j.apenergy.2020.116239.

Mayer, M. J., and D. Yang, 2022: Probabilistic photovoltaic power forecasting using a calibrated ensemble of model chains. *Renewable and Sustainable Energy Reviews*, **168**, 112821, https://doi.org/10.1016/j.rser.2022.112821.

Mayer, M. J., and D. Yang, 2023: Pairing ensemble numerical weather prediction with ensemble physical model chain for probabilistic photovoltaic power forecasting. *Renewable and Sustainable Energy Reviews*, **175**, 113171, https://doi.org/10.1016/j.rser.2023.113171.

Messner, J. W., G. J. Mayr, A. Zeileis, and D. S. Wilks, 2014: Heteroscedastic extended logistic regression for postprocessing of ensemble guidance. *Mon. Wea. Rev.*, **142**, 448–456, https://doi.org/10.1175/MWR-D-13-00271.1.

Mikolov, T., K. Chen, G. Corrado, and J. Dean, 2013: Efficient estimation of word representations in vector space. *Proc. 1st International Conf. on Learning Representations*, Scottsdale, Arizona, USA.

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

Rasp, S., and S. Lerch, 2018: Neural networks for postprocessing ensemble weather forecasts. *Mon. Wea. Rev.*, **146**, 3885–3900, https://doi.org/10.1175/MWR-D-18-0187.1.

Reda, I., and A. Andreas, 2004: Solar position algorithm for solar radiation applications. *Solar Energy*, **76**, 577–589, https://doi.org/10.1016/j.solener.2003.12.003.

Reindl, D. T., W. A. Beckman, and J. A. Duffie, 1990: Evaluation of hourly tilted surface radiation models. *Solar Energy*, **45**, 9–17, https://doi.org/10.1016/0038-092X(90)90061-G.

Roberts, J. J., A. A. Mendiburu Zevallos, and A. M. Cassula, 2017: Assessment of photovoltaic performance models for system simulation. *Renewable and Sustainable Energy Reviews*, **72**, 1104–1123, https://doi.org/10.1016/j.rser.2016.10.022.

Scheuerer, M., 2014: Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quart. J. Roy. Meteor. Soc.*, **140**, 1086–1096, https://doi.org/10.1002/qj.2183.

Scheuerer, M., M. B. Switanek, R. P. Worsnop, and T. M. Hamill, 2020: Using artificial neural networks for generating probabilistic subseasonal precipitation forecasts over California. *Mon. Wea. Rev.*, **148**, 3489–3506, https://doi.org/10.1175/MWR-D-20-0096.1.

Schulz, B., and S. Lerch, 2022: Machine learning methods for post-processing ensemble forecasts of wind gusts: A systematic comparison. *Mon. Wea. Rev.*, **150**, 235–257, https://doi.org/10.1175/MWR-D-21-0150.1.

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

Sengupta, M., Y. Xie, A. Lopez, A. Habte, G. Maclaurin, and J. Shelby, 2018: The national solar radiation data base (NSRDB). *Renewable and Sustainable Energy Reviews*, **89**, 51–60, https://doi.org/10.1016/j.rser.2018.03.003.

Song, M., and Coauthors, 2024: Non-crossing quantile regression neural network as a calibration tool for ensemble weather forecasts. *Adv. Atmos. Sci.*, **41**, 1417–1437, https://doi.org/10.1007/s00376-023-3184-5.

Taillardat, M., O. Mestre, M. Zamo, and P. Naveau, 2016: Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Mon. Wea. Rev.*, **144**,

2375–2393, https://doi.org/10.1175/MWR-D-15-0260.1.

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

Thorarinsdottir, T. L., and T. Gneiting, 2010: Probabilistic forecasts of wind speed: Ensemble model output statistics by using heteroscedastic censored regression. *Journal of the Royal Statistical Society Series A: Statistics in Society*, **173**, 371–388, https://doi.org/10.1111/j.1467-985X.2009.00616.x.

van der Meer, D. W., J. Widén, and J. Munkhammar, 2018: Review on probabilistic forecasting of photovoltaic power production and electricity consumption. *Renewable and Sustainable Energy Reviews*, **81**, 1484–1512, https://doi.org/10.1016/j.rser.2017.05.212.

Vannitsem, S., and Coauthors, 2021: Statistical postprocessing for weather forecasts: Review, challenges, and avenues in a big data world. *Bull. Amer. Meteor. Soc.*, **102**, E681–E699, https://doi.org/10.1175/BAMS-D-19-0308.1.

Veldkamp, S., K. Whan, S. Dirksen, and M. Schmeits, 2021: Statistical postprocessing of wind speed forecasts using convolutional neural networks. *Mon. Wea. Rev.*, **149**, 1141–1152, https://doi.org/10.1175/MWR-D-20-0219.1.

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

Yagli, G. M., D. Yang, and D. Srinivasan, 2020: Ensemble solar forecasting using data-driven models with probabilistic post-processing through GAMLSS. *Solar Energy*, **208**, 612–622, https://doi.org/10.1016/j.solener.2020.07.040.

Yang, D., 2019: A guideline to solar forecasting research practice: Reproducible, operational, probabilistic or physically-based, ensemble, and skill (ROPES). *Journal of Renewable and Sustainable Energy*, **11**, 022701, https://doi.org/10.1063/1.5087462.

Yang, D., 2020: Ensemble model output statistics as a probabilistic site-adaptation tool for solar irradiance: A revisit. *Journal of Renewable and Sustainable Energy*, **12**, 036101, https://doi.org/10.1063/5.0010003.

Yang, D., and C. A. Gueymard, 2020: Ensemble model output statistics for the separation of direct and diffuse components from 1-min global irradiance. *Solar Energy*, **208**, 591–603, https://doi.org/10.1016/j.solener.2020.05.082.

Yang, D., and D. van der Meer, 2021: Post-processing in solar forecasting: Ten overarching thinking tools. *Renewable and Sustainable Energy Reviews*, **140**, 110735, https://doi.org/10.1016/j.rser.2021.110735.

Yang, D., E. Wu, and J. Kleissl, 2019: Operational solar forecasting for the real-time market. *International Journal of Forecasting*, **35**, 1499–1519, https://doi.org/10.1016/j.ijforecast.2019.03.009.

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

Zhong, X., L. Chen, X. Fan, W. Qian, J. Liu, and H. Li, 2024: FuXi-2.0: Advancing Machine Learning Weather Forecasting Model for Practical Applications. Available from https://arxiv.org/abs/2409.07188 .