Solar Energy Advances 4 (2024) 100059

Contents lists available at ScienceDirect

Solar Energy Advances

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

Enhancing the reliability of probabilistic PV power forecasts using conformal prediction

Yvet Renkema [a], Lennard Visser [b], Tarek AlSkaif [a,*]

[a] *Information Technology Group, Wageningen University, Wageningen, The Netherlands*
[b] *Copernicus Institute of Sustainable Development, Utrecht University, Utrecht, The Netherlands*

ARTICLE INFO

Keywords:
Conformal prediction
Machine learning
Photovoltaic power
Regression methods
Solar power forecasting

ABSTRACT

The increasing integration of renewable energy, particularly solar photovoltaic (PV) power, presents challenges for power system operation. Accurate forecasts of renewable energy are both financially beneficial for electricity suppliers and necessary for grid operators to optimize operation and avoid grid imbalances. This paper proposes a forecasting framework to implement conformal prediction (CP) on top of point prediction models, which predict the PV power on a day-ahead basis, to quantify the uncertainty of those predictions. Simple and multiple linear regression, along with random forest regression, are used to construct the point predictions based on weather forecasts. Several variants of CP, including weighted CP, CP with k-nearest neighbors (KNN), CP with Mondrian binning, and conformal predictive systems, are built to transform the point predictions into rigorous uncertainty intervals or cumulative distribution functions to enhance reliability. The framework's performance is evaluated using large datasets of weather predictions and PV power output in the Netherlands. Results indicate that CP combined with KNN and/or Mondrian binning after a linear regressor outperforms the corresponding linear quantile regressor. CP with KNN and Mondrian binning after using random forest regression demonstrates the most accurate probabilistic PV power forecasts, improving the weighted interval score by 14% compared to multiple linear quantile regression.

1. Introduction

Solar photovoltaic (PV) systems have seen an exponential growth since the beginning of the century, triggered by supporting climate policies and rapidly decreasing costs. Total solar PV power is expected to increase on average by 13% per year from 2020 to 2030, raising the combined share of PV and wind power in the global electricity generation from 9% in 2019 to 30% in 2030 [1]. These sustainable energy sources are free of direct carbon emissions. However, the integration of solar PV and wind energy in the power system poses serious operational challenges due to their intermittency, non-dispatchability and unpredictability [2]. This increases the overall uncertainty in the power system operation and therefore the need for reliable predictions of the PV power output.

Accurate PV power forecasting is essential for numerous decision-making processes within power systems, such as reducing operating reserve capacity, generating precise bids in electricity markets, and maintaining grid stability [3,4]. This can potentially reduce integration costs associated with high PV penetration. Two main forecasting approaches exist: point (or single-value) forecast and probabilistic forecast. In point prediction approaches, one value is predicted for each time on the horizon. Existing literature reviews highlight extensive research on solar irradiance and PV power forecasting [5–7], while also emphasizing the necessity for reliable and accurate PV predictions tailored to power system requirements, and for comparative analysis among different point prediction models [8]. Probabilistic forecasting approaches enhance the reliability of predictions by providing information about their full probability distributions, thereby enhancing informativeness, which is crucial for decision-making under risk, such as bidding in electricity markets. A review of common probabilistic methods used for PV power forecasting is provided in [9], while [10] presents a comparative analysis of these methodologies and the factors affecting their accuracy.

Conformal prediction (CP) is an emerging probabilistic forecasting method [11]. In its most basic form, the residuals of predictions from a calibration dataset are used to calibrate prediction intervals from a test dataset. CP offers a measure of confidence or credibility by transforming point predictions into prediction intervals with a probabilistic guarantee of covering the true outcome [11]. This makes CP particularly useful in situations with uncertainty where reliability is essential, such as in decision-making processes. Moreover, CP is

[*] Corresponding author.
E-mail address: tarek.alskaif@wur.nl (T. AlSkaif).

https://doi.org/10.1016/j.seja.2024.100059
Received 6 December 2023; Received in revised form 29 March 2024; Accepted 5 April 2024
Available online 15 April 2024

distribution-free and model-agnostic, meaning that it can be combined with any point prediction model [11]. CP guarantees a user-specified probability that on average the prediction interval contains the correct value. This property is called marginal coverage. Conditional coverage is a stronger property guaranteeing that for every test value, CP returns an interval with the user-specified probability. Various adaptations of CP contribute to conditional guarantees, making it a very flexible and promising probabilistic forecasting method. Seminal research has been done using CP on time-series data, demonstrating the promising performance of this emerging probabilistic forecasting method [12–14]. However, time-series data is the broader category to which PV power data belongs and therefore, current research on time-series data does not account for the specific application requirements for PV output power forecasting. The authors of [15,16] have used CP models to predict wind power intervals. In [17], a CP model is used for day-ahead energy demand forecasting. In these three papers, the proposed CP-based models outperformed the benchmarks on accuracy or on interval width and coverage. However, studies on using CP to quantify the uncertainty of PV power predictions are still lacking and is, therefore, one of the main novelties of this research.

This paper proposes and investigates the added value of a framework using a wide variety of CP methods to enhance the reliability of day-ahead solar PV power forecasting. This framework incorporates simple and multiple linear regression (SLR and MLR) as well as random forest regression (RFR) as point prediction models that predict the day-ahead electricity supply of PV systems based on weather forecasts. The residuals of the point predictions on the calibration dataset are used in various variants of CP. This process results in calibrated prediction intervals or cumulative distribution functions (CDFs) for the test dataset, providing a quantification of the uncertainty associated with point predictions. These prediction intervals and CDFs offer a probabilistic guarantee of encompassing the true outcome. Finally, the performance of this framework is assessed using weather predictions and PV power measurements from the Netherlands. This research holds particular relevance for electricity market participants seeking to maximize profit while managing associated risks. Additionally, it offers valuable insights for grid operators, aiding in the anticipation and mitigation of expected grid imbalances. The main contributions of this article can be summarized as:

- A novel framework using CP to enhance the reliability of probabilistic day-ahead PV power forecasting.
- Developing and applying multiple CP variants to point prediction models for quantifying uncertainty.
- Evaluating the performance and benchmarking the CP variants using actual weather predictions and PV power data from the Netherlands.

The structure of the paper is as follows. Section 2 provides a description of the machine learning-based regression models, the uncertainty quantification with linear quantile regression (LQR) and the CP methods. Section 3 presents a performance evaluation of the point prediction models and the uncertainty quantification methods. Finally, the paper is concluded in Section 4 which also provides pointers for future work.

2. Methods

Regression methods are commonly used in solar power forecasting applications due to their ability to model the relationship between solar irradiance, weather variables, and PV power output. Additionally, the regression methods considered in this study offer simplicity, flexibility, and interpretability, making them suitable for capturing the complex dynamics of solar PV power generation. Recently, deep learning models have received increasing attention, and although it is deemed very efficient for image and language learning, this does not necessarily hold for tabular data. Multiple research projects show that RFR performs at least as good as long-short-term-memory (LSTM) networks for forecasting of electricity consumption or solar PV power [18,19]. For tabular data, deep learning models are prone to be too sensitive to uninformative features and too smooth while tree-based models consider these irregular patterns and are more robust [20]. Therefore, simple machine learning models should be considered for time-series forecasting [18,20,21]. For this study, commonly known machine learning models (e.g., regression methods) are used to generate point predictions of the PV power output.

These point prediction models have some shortcomings in the considered PV power forecasting application. For instance, linear regression (LR) assumes a normal conditional distribution and a constant variance for the response variable. Both assumptions do not apply to the PV data. Additionally, LR only describes the relationship between the independent variables and the mean of the response variable. On the other hand, LQR points out relationships between a specific quantile of the response variable and the independent variable(s), producing probabilistic forecasts. Hence, LQR preserves more information about the full conditional distribution of the response variable compared to LR.

Machine learning models such as RFR fail to properly estimate the uncertainty of their predictions [22] and most quantile methods like quantile regression forests do not provide probabilistic guarantees, which is where CP comes into play. CP is a relatively new framework with an increasing amount of publications on the subject each year. In Scopus, it has been rising from no publications in 2006 to 30 in 2015 and 73 in 2022.[1] There are both conformal regressors and conformal classifiers. Conformal regressors transform point predictions into uncertainty intervals without the need for distributional assumptions on the data [11]. Those uncertainty intervals are rigorous, indicating that they have a probabilistic guarantee of covering the true outcome. In other words, CP guarantees marginal coverage, for which the user chooses the error rate, α. This study focuses on regression, and therefore any mention of CP refers to conformal regressors.

The methods and steps followed in the paper are summarized in Fig. 1. Section 2.1 starts with an explanation of the point prediction models that are used which is followed by Section 2.2 on the uncertainty quantification methods. The main uncertainty quantification methods are presented by the various CP variants and they are benchmarked against LQR methods.

2.1. Regression methods

Besides an RFR model, SLR and MLR are used in this research, see Fig. 1. MLR is a simple yet effective regression model that is widely adopted to forecast or estimate solar PV power [8,23]. Based on training data, the MLR model uses a loss function to determine the coefficients that explain a linear relation between the predictor variables and the target variable.

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \cdots + \cdots + \beta_k \mathbf{x}_k, \tag{1}$$

where y is the target variable (i.e., solar PV power), βs are the regression coefficients, $x_1, x_2, \dots, x_k$ are the predictor variables and k is the number of predictor features. An SLR model is similar to a MLR model with $k = 1$.

RFR is a tree-based regression model that has proven its value for time-series forecasting and regression applications [18,20]. It will also be used in this research to forecast solar PV power and to compare its performance with SLR and MLR. RFR is expected to outperform the LR models as PV power data shows nonlinear relationships with its

[1] Using scope *title, abstract and keyword* on the 6th of February 2023 with the search query: (“conformal predict*” OR “conformal inference” OR “conformal regressor”).

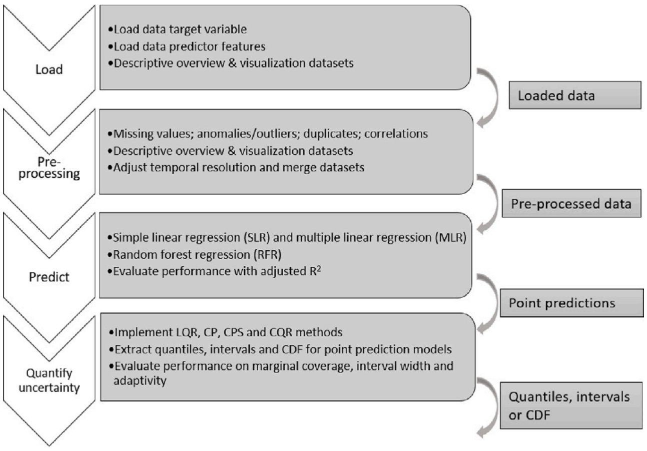
Load
•Load data target variable
•Load data predictor features
•Descriptive overview & visualization datasets
Loaded data
Pre-processing
•Missing values; anomalies/outliers; duplicates; correlations
•Descriptive overview & visualization datasets
•Adjust temporal resolution and merge datasets
Pre-processed data
Predict
•Simple linear regression (SLR) and multiple linear regression (MLR)
•Random forest regression (RFR)
•Evaluate performance with adjusted R²
Point predictions
Quantify uncertainty
•Implement LQR, CP, CPS and CQR methods
•Extract quantiles, intervals and CDF for point prediction models
•Evaluate performance on marginal coverage, interval width and adaptivity
Quantiles, intervals or CDF

Fig. 1. A schematic of the general flow of the data with its form after each step depicted on the right.

predictors [24], which LR cannot take into account [25]. Moreover, RFR tends to yield more efficient conformal predictors compared to other models [26]. RFR is an ensemble based model that consists of a number of trees, each made up of n layers and 2^n decision nodes, with $n = 0$ at the first layer. The decision trees are created independently and are built by considering bootstrap samples of the training dataset. Next, for each tree a random subset of the predictor features is considered to construct the decision nodes by optimizing on a loss function, e.g., least squares [27]. The output of an RFR model is equal to the conditional mean of all constructed trees.

For SLR the predictor feature is surface solar radiation downwards (SSRD), as it has the highest correlation with PV power. The model, fitted with ordinary least squares, indeed shows a p-value of zero for SSRD and, therefore, points out a significant relationship between SSRD and PV power. To select the optimal predictor features for MLR, forward subset selection is used in this research with 10-fold cross-validation. The selected predictor features are time horizon, zonal wind speed, total cloud cover, surface solar radiation (SSR) and the cosine of the hour of the day (HoD). The variance inflation factor is calculated showing no problematic collinearity between the selected features. The hyperparameter tuning for RFR indicates that the model performs best with 375 trees and by considering three features when looking for the best split. The model is fitted with those hyperparameters and the test datapoints are predicted with the fitted model. A couple of studies compare the point prediction models preceding CP and conclude that RFR yields more efficient conformal predictors compared to, among others, neural networks, KNN and gradient boosted tree models [26,28].

2.2. Uncertainty quantification methods

For the quantification of uncertainty, simple linear quantile regression (SLQR) and multiple linear quantile regression (MLQR) are used as benchmark models. Their predictive features are similar to those for SLR and MLR. Basic CP and several variants on CP are used as well and are explained in the sections below. A summary of the CP variants used can be found in Table 1.

2.2.1. Basic CP

As a first step of the most basic variant for CP, a point prediction model is used to predict on the calibration dataset after which the residuals of these predictions are extracted. The absolute values of the residuals are then called the nonconformity scores, indicating how ‘atypical’ a certain datapoint is. Secondly, the variable q̂ is defined based on the chosen value of α and the sorted nonconformity scores such that a fraction α of the calibration datapoints have nonconformity scores exceeding q̂ (see Fig. 2, where $\hat{q}$ is calculated as the percentile corresponding to $(1 - \alpha)$. Lastly, the point prediction model is used to predict the test data and q̂ is both added and subtracted from the point predictions to derive the prediction intervals. This is shown in Eq. (2), where PI_i stands for the prediction interval and pp_i is the point prediction of test point i.

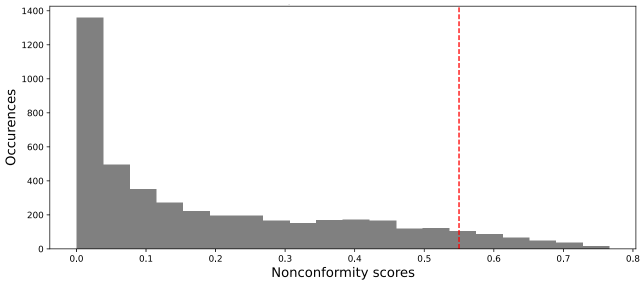

Fig. 2. Example of how q̂ is extracted from the sorted nonconformity scores, where α is chosen to be 0.10 following that $\hat{q}$ is calculated as the percentile corresponding to $(1 - \alpha)$. The red line shows the value for q̂, in this case 0.55, where a fraction α of the nonconformity scores exceeds q̂.

$$PI_i = [pp_i - \hat{q}, pp_i + \hat{q}]. \tag{2}$$

As previously mentioned, marginal coverage is guaranteed with CP, i.e. on average the chosen error rate, α, is realized. Satisfying the error rate for each type of datapoint is called conditional coverage [11]. The property of a method to give wider intervals for points that are harder to predict than for ‘easy’ points, is called adaptivity. Most variants on basic CP aim to increase adaptivity to approximate conditional coverage. However, conditional coverage can, in most cases, only be approached instead of fully achieved [11]. For this study, the basic CP is extended with multiple variations which is described in the following sections.

2.2.2. Weighted CP

To consider the distribution drift of time-series data, weighted CP can be used implying increased pre-defined weights are given to nonconformity scores of points closer to the test datapoint [11]. For this study, a sliding window of k preceding points in the test dataset with their predicted and actual values are used as the calibration dataset for a basic method for weighted CP. To account for the day-ahead forecast horizon, the window of the k points to be used for calibration is shifted so that the 24 h before the considered test datapoint are not used for calibrating the prediction interval of that test datapoint.

Additionally, distance-related weighted CP is adopted in this study, where the k preceding points are given weights according to their time-based distance to the test point with increased weights for points closer-by. The distance-related weights are linearly increasing such that they add up to one which is shown in Eq. (3). In the equation $k = 1$ is the furthest point away from the considered test point, and k is the closest point in time to the considered test point.

$$weight_k = \frac{k}{\sum_{i=1}^{k} i} = \frac{k}{\frac{k(k+1)}{2}} = \frac{2}{k+1}. \tag{3}$$

Moreover, hour-related weighted CP is used, where in addition to the basic weighted CP, only datapoints in the window of size k with a similar hour of the day as the test point are considered for the calibration dataset. ‘Similar’ is defined as not deviating more than one hour from the hour of the considered test datapoint. Also, weighted CP with both distance- and hour-related weights is evaluated where linearly increasing weights are given to points closer-by the test datapoint, but only if they have a similar hour of the day as the test datapoint.

2.2.3. CP with uncertainty scalars

The most basic version of CP results in an interval width of $2 * \hat{q}$ for all predictions, yet certain points are harder to estimate accurately than others. In CP with uncertainty scalars, the so-called difficulty estimates of the point predictions are implemented to adjust the interval sizes [29]. Points with high difficulty estimate are expected to be hard to estimate and thus more uncertain and, therefore, yield wider prediction intervals with this CP variant. The nonconformity scores are then formulated as the absolute values of the residuals from the calibration dataset divided by their difficulty estimates (see Eq. (4), where NS_j is the nonconformity score for calibration point j and de_j the difficulty estimate belonging to calibration point j).

$$NS_j = \frac{|\, residual_j \,|}{de_j}. \tag{4}$$

Then, using those nonconformity scores q̂ is determined just as in the basic CP. The prediction interval is in this case the difficulty estimate of the test point multiplied by q̂ added and subtracted from the point prediction (see Eq. (5), where PI_i and pp_i stand for the prediction interval and point prediction, respectively, of test point i, and de_i is the difficulty estimate of test point i).

$$PI_i = [pp_i - de_i * \hat{q}, pp_i + de_i * \hat{q}]. \tag{5}$$

For this study, two difficulty estimates, or uncertainty scalars, are explored. Firstly, using the predicted residuals for a test point. To predict the residuals a regression model is trained on the residuals from the train dataset. This prediction model has the same characteristics as the corresponding point prediction model used for predicting the PV power. The second type of difficulty estimates is derived by taking the average residuals of the k-nearest neighbors (KNN), the points in the calibration dataset that are most similar with respect to the independent variables. In an iterative process, the optimal value for the parameter k is determined to be 50. A lower value for the number of neighbors leads to overfitting, while a higher value reduces the adaptivity.

2.2.4. CP with Mondrian binning

CP can also be performed after splitting both the calibration and the test dataset into Mondrian categories, which is called Mondrian binning or simply binning. In Mondrian binning, a predefined number of equal-sized bins are created from the calibration dataset based on the predicted values for PV power. The threshold values for these bins are extracted and applied to the test dataset. CP is then performed for each bin separately resulting in a value for $\hat{q}_b$ for each bin b which is then applied to the test datapoints belonging to that bin. This is shown in Eq. (6), where PI_i and pp_i stand for the prediction interval and point prediction, respectively, of test point i, and $\hat{q}_b$ is the specific $\hat{q}$ for each bin b. In previous research, CP with bins was found to outperform basic CP by differing the interval widths between bins and thus creating adaptivity [30]. Based on empirical evaluation the optimal number of bins in this study is found to be 15, as more than 15 bins leads to overfitting, while fewer bins reduce the adaptivity.

$$PI_i = [pp_i - \hat{q}_b, pp_i + \hat{q}_b]. \tag{6}$$

2.2.5. CPS

An upcoming CP variant is conformal predictive systems (CPS) which outputs conformal predictive distributions (CPD), i.e. CDFs. CPS uses the residuals instead of the absolute values of the residuals as nonconformity scores [31]. Consequently, a prediction is not necessarily centered in the middle of an interval. In other words, the intervals can be 'shifted' and the left and right hand side of the intervals are not by definition equal. Therefore, CPS is more flexible than CP and thus preserves more information. Most of the variants for CP can also be applied to CPS. In this study, both KNN and binning are used in combination with CPS. Methods with either CPS, KNN and/or binning are applied with the help of the crepes package (version 0.1.0). For the other variants of CP no packages have been used, instead the functions were built by the authors.

Table 1
Summary of the CP methods used in this study with their abbreviations.

Abbreviation	Method
M1	Basic CP
M2	Weighted CP
M3	Distance-related weighted CP
M4	Hour-related weighted CP
M5	Distance- and hour-related weighted CP, a combination of M3 and M4.
M6	CP with the predicted residuals as an uncertainty scalar
M7	CP with KNN as an uncertainty scalar
M8	CP with Mondrian binning
M9	CP with KNN as an uncertainty scalar and with Mondrian binning
M10	Basic CPS
M11	CPS with KNN as an uncertainty scalar
M12	CPS with Mondrian binning
M13	CPS with KNN as an uncertainty scalar and with Mondrian binning
M14	CQR to conformalize intervals from SLQR and MLQR.

2.2.6. CQR

Conformalized quantile regression (CQR) uses a quantile regression algorithm producing predictions for conditional quantiles as a preceding model [11]. Similarly, a model outputting predictions for specific confidence intervals can be used. Seeing that those predicted quantile values or confidence intervals are already adaptive to the predictive features, CQR has some inherited adaptivity as well. A nonconformity score for CQR is the difference between the actual value for a calibration point and its nearest predicted interval margin which can be either the lower or the upper margin [11]. This difference is negative if the actual value lies inside of the interval and positive if it is outside of the interval. Similarly to the basic variant of CP, these nonconformity scores are sorted in increasing order and q̂ is defined based on α. Subsequently, the prediction intervals from the test dataset are extended or reduced when considered too confident or conservative, respectively [11]. This is shown in Eq. (7), where PI_i stands for the prediction interval of test point i, and lb_i and ub_i are the lower and upper bounds for test point i as predicted by the quantile regressor. If, for instance, the error rate is lower than α when applying only a quantile regression model on the calibration dataset, q̂ will be negative. Then the intervals are considered too conservative and the intervals from the test dataset are contracted.

$$PI_i = [lb_i - \hat{q}, ub_i + \hat{q}]. \tag{7}$$

3. Results and discussions

3.1. Datasets

The target data originate from an open-source dataset with power measurements of 175 PV systems in the province of Utrecht, the Netherlands [32]. The power measurements have a one minute resolution and cover January 2014 until December 2017. A quality control routine with single and across system filters is already applied to the dataset in the form of a Python package that is publicly available [32]. Partly due to the quality control routine, this dataset contains many NaN values. On average 29% of the data is missing, ranging from 12% to 77% for the single PV systems. Therefore, in this research, the 175 PV systems are aggregated as shown in Fig. 3. However, before this aggregation, the values are normalized per PV system to level out differences in sizes and to get values between zero and one. The resolution is converted

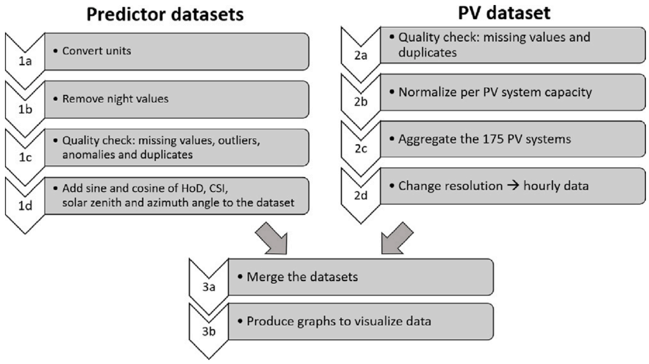
Predictor datasets
1a
• Convert units
1b
• Remove night values
1c
• Quality check: missing values, outliers, anomalies and duplicates
1d
• Add sine and cosine of HoD, CSI, solar zenith and azimuth angle to the dataset
PV dataset
2a
• Quality check: missing values and duplicates
2b
• Normalize per PV system capacity
2c
• Aggregate the 175 PV systems
2d
• Change resolution → hourly data
3a
• Merge the datasets
3b
• Produce graphs to visualize data

Fig. 3. Flow diagram of the pre-processing steps of the PV and the predictor datasets. CSI stands for clear sky irradiance and HoD for hour of the day.

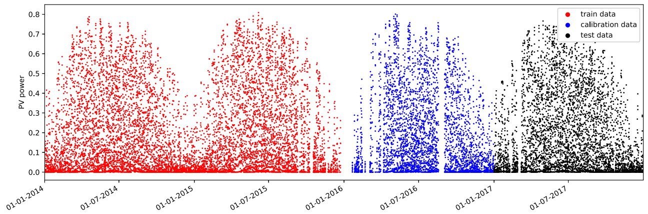

Fig. 4. The normalized PV power values over time split into train, calibration and test data. The 1st of January of 2016 and 2017 are used as breaking points to split the dataset.

to hourly by taking the mean over the values in each hour. Python 3.8.10 is used for processing the data. A visual inspection of the hourly aggregated data does not indicate outliers.

The dataset used for predictor data comes from the Meteorological Archival and Retrieval System of the European Centre for Medium-Range Weather Forecasts (ECMWF) [33]. This dataset contains hourly weather predictions for January 2014 till December 2017 on noon of day T for day T+1, matching the requirements of the day-ahead solar PV forecasting (e.g., 24 h ahead with a 1 h resolution). It contains predictions on variables such as surface pressure, cloud cover, wind speed, temperature, precipitation and solar radiance. The predictions are for *De Bilt*, which is located in the province of Utrecht. First, some units have been converted and the night values for SSR and SSRD have been set to zero. Thereafter, the dataset contains no missing values, outliers or anomalies.

Research has shown that machine learning models improve by extending the dataset with physical relationships between PV power and the environment [19,34]. Therefore, the data is complemented with the cosine of the HoD to reflect its cyclic nature [23]. Additionally, the clear sky irradiance, solar zenith and azimuth angle are included with the help of the PVlib package as they reflect the position of the sun over time [35]. The pre-processing of the datasets is summarized in Fig. 3.

The dataset is split by date with years 2014 and 2015 as train dataset, 2016 as calibration dataset and 2017 as test dataset as shown in Fig. 4. Each year consists of around four to five thousand day value datapoints. The night values are always zero, therefore, the model is trained, calibrated and tested on the day values only.

3.2. Regression methods

Fig. 5 visualizes the point predictions from SLR, MLR and RFR on the test dataset. Table 2 shows the adjusted R^2 and RMSE for each point prediction model on the test data. The RFR performs best due to its ability to handle nonlinear relationships, followed by MLR and SLR in that order.

Table 2
Values for the error metrics per point prediction model with the best value in bold.

Model	Adjusted R^2	RMSE
SLR	0.806	0.087
MLR	0.821	0.083
RFR	**0.854**	**0.075**

Table 3
The error metrics for the 90% confidence intervals with the abbreviations between brackets showing what the metrics are an indicator of, namely marginal coverage (MC), interval width (IW), and/or adaptivity (AD).

Evaluation metric	Description
Breach	Takes a value of 0 if the coverage is above 0.90, meaning that between 90% and 100% of the test datapoints fall within their predicted intervals. Otherwise, the breach is 0.90–the coverage. For instance, if the coverage is 0.83, the breach is 0.07. A lower value is preferred for this metric (MC)
Sharpness	Average interval width (IW)
Calibration	The sum of the penalties for predictions outside of the intervals. A penalty is the distance to the nearest interval margin multiplied by two divided by α. Therefore, for smaller values of α, the penalty is more severe (MC)
Interval score	The sum of the sharpness and calibration (IW + MC)
SSC	The lowest average coverage of all width bins (AD)

3.3. Uncertainty quantification methods

All LQR and CP models are evaluated on marginal coverage, interval width, and adaptivity with the help of a few error metrics. The interval score is an error metric consisting of two elements, namely sharpness and calibration. Sharpness is the average interval width and calibration is the sum of the penalties for test points outside of the interval which become more severe for decreasing values of α [36]. To evaluate adaptivity of the CP methods, the size-stratified coverage (SSC) metric is used [11]. For SSC, each test datapoint is placed in a bin based on the interval width after which for each bin the coverage is calculated. In an ideal situation, the coverage rate is equal to 1-α for each bin seeing that the bins represent the difficulty of estimating the point. To get the SSC, the lowest coverage rate is then extracted.

First, a few error metrics are collected for $\alpha = 0.10$, so for the 90% confidence intervals. Thereafter, by taking α between 0.02 and 0.98 with steps of 0.02, confidence intervals for 49 confidence levels are gathered. A summary of the evaluation metrics considered for the 90% confidence intervals in this study is provided in Table 3.

The scores for breach and SSC for MLQR and the CP methods combined with RFR are shown in Table 4 and those results along with the results of MLQR and the CP methods combined with SLR and MLR are discussed in the following paragraphs. First, all results from the 90% confidence intervals are discussed.

For CQR, $\hat{q}$ is consistently zero, therefore it can be concluded that CQR (M14) does not change nor improve the results from the LQR models. LQR models are already adaptive to their predictive features. This automatically results in adaptivity for CQR, making it an interesting method to explore further. The current study only touches upon the most basic form of CQR while additions like weights or binning are expected to improve CQR.

Based on the size of the test dataset (4514 datapoints) and α (0.10), a formula proposed by [11] shows that a coverage of less than 89.2% (ϵ > 0.008) is the threshold value that is required for the CP methods to be valid. For all weighted CP methods (M2 till M5) with a value of k below 100, an unacceptable breach is detected, indicating an insufficient calibration set size. On the contrary, the following uncertainty quantification methods result in marginal coverage over 90% for all point prediction models: LQR, hour-related weights with K=300,

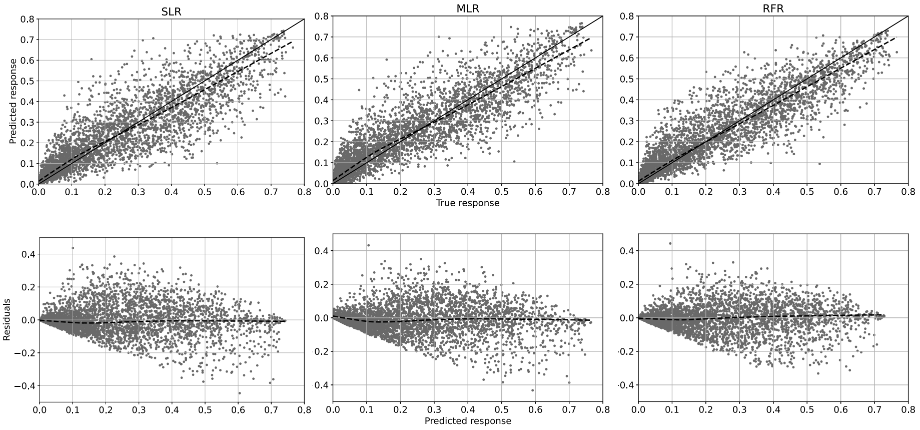

Fig. 5. Upper scatter plots depict predicted against true response with a trend line (dashed) and ideal fit (solid) on the test dataset for SLR, MLR and RFR. The lower scatter plots show the corresponding residuals against the true response on the test dataset.

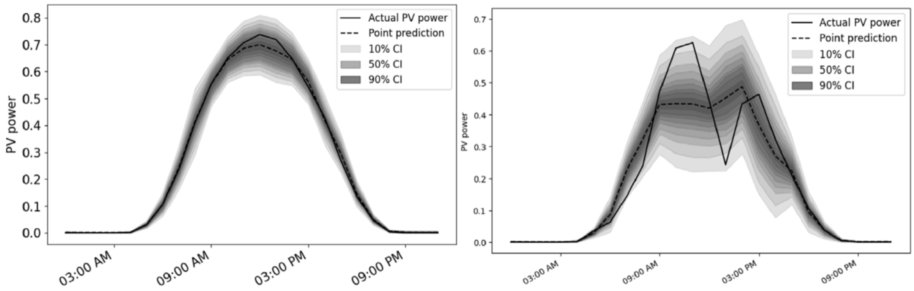

Fig. 6. Normalized point predictions from the RFR model and confidence intervals from M9 after RFR for two days: a clear sky day (left) on June 1st 2017, and a day with cloud cover (right) on June 5th 2017.

Table 4
Table showing the results gathered from the 90% confidence intervals and the 49 combined confidence levels for MLQR and RFR with M1 till M14 (see Table 1). M2 till M5 are all for K=300. Lower values are preferred on all metrics, except SSC where a high value indicates good performance. Cali. stands for calibration. Best value for WIS is shown in bold.

Method	Version	90% confidence intervals		49 confidence levels		
		Breach	SSC	WIS	Sharpness	Cali.
LQR	MLQR	0.000	0.854	0.163	0.067	0.096
M1	RFR	0.000	0.825	0.152	0.040	0.111
M2	RFR	0.005	0.840	0.155	0.039	0.116
M3	RFR	0.005	0.843	0.155	0.039	0.116
M4	RFR	0.000	0.865	0.145	0.054	0.091
M5	RFR	0.000	0.869	0.145	0.055	0.090
M6	RFR	0.000	0.970	0.142	0.077	0.065
M7	RFR	0.000	0.808	0.142	0.044	0.098
M8	RFR	0.013	0.855	0.143	0.059	0.084
M9	RFR	0.003	0.863	**0.140**	0.061	0.079
M10	RFR	0.000	0.822	0.154	0.042	0.112
M11	RFR	0.002	0.802	0.143	0.045	0.098
M12	RFR	0.015	0.842	0.144	0.056	0.088
M13	RFR	0.000	0.876	0.142	0.057	0.085
M14	MLQR	0.000	0.854	0.163	0.067	0.096

predicted residuals as uncertainty scalars, KNN and/or binning after an LR point prediction model and CQR.

Furthermore, it is shown that CP after MLR with hour-related weights for K=300, CP with residuals as an uncertainty scalar after RFR and CP and CPS with binning after MLR score best on adaptivity. However, the former produces very wide intervals. Besides, it is concluded CPS methods have slightly better values for the SSC than their corresponding CP methods (comparing M1 with M10, M7 with M11, M8 with M12 and M9 with M13).

As a consequence of the high accuracy of RFR as stated in Section 3.2, the CP methods based on this point prediction model produce the smallest intervals. Over all point prediction models, the method with KNN (M7) and the method with distance- and hour-related weights (M5) give the smallest intervals while LQR, the predicted residuals as an uncertainty scalar (M6) and CQR (M14) result in wider intervals. The latter methods do yield high marginal coverage as concluded earlier in this section which could be explained by the wide intervals.

Generally, methods with a high breach have a high value for calibration. Where the breach simply considers if a point falls in- or outside the interval, calibration also penalizes more severe if the point is further away from the interval. For the method with binning after RFR (M8), the breach is quite high while the calibration is low. This indicates a relatively large amount of points outside the interval, but that they are quite close to the bounds of the interval.

By taking α between 0.02 and 0.98 with steps of 0.02, confidence intervals for 49 confidence levels are created. Fig. 6 shows an example of confidence intervals predicted by M9 for two days with different weather conditions in June 2017. Having multiple confidence intervals for each datapoint allows for calculating the weighted interval score (WIS) and its partials, sharpness and calibration. This is an elaborate version of the interval score and is proven to approximate the continuous ranked probability score (CRPS) [36]. Although both CRPS and WIS are used to evaluate the performance of probabilistic forecasts, WIS uniquely assesses prediction intervals. CRPS, on the other hand, is based on CDFs, and while the CPS method generates CDFs, the majority of the other considered CP methods output prediction intervals. This is why WIS was preferred over CRPS in this study.

Fig. 7 shows the WIS for all LQR methods and CP methods in combination with the point prediction models. It shows that for both LR models, the best performing methods based on WIS are, in decreasing order of performance: CP with KNN and/or bins (M7 till M9), LQR, CQR

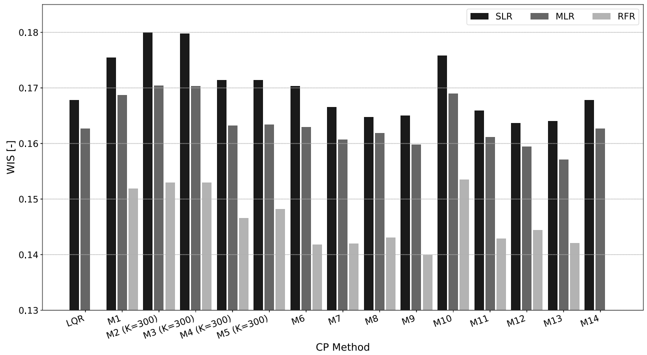

Fig. 7. Bar chart showing the WIS for several methods in combination with one of the three point prediction models. M1 till M14 refer to the list of methods in Table 1.

(M14), the method with predicted residuals as an uncertainty scalar (M6) and the method with hour-related weights (M4). Methods with RFR as a point prediction model always outperform methods with LR as point prediction models with respect to the WIS (as shown in Fig. 7).

The WIS and its partials for MLQR and for the CP methods combined with RFR are shown in Table 4. A compromise is made on sharpness to obtain a low WIS seeing that methods with low WIS values have relatively high proportions of sharpness, while the ones with higher WIS values have lower proportions for sharpness. CP with KNN and bins after RFR (as shown in Table 4 under M9) is superior to all other methods based on WIS followed by CP with predicted residuals as an uncertainty scalar after RFR (M6), but they both lack sharpness. CP with KNN after RFR (M7) has the third lowest WIS and scores better on sharpness. The WIS values of comparable CP and CPS (i.e. M7 with M11, M8 with M12, M9 with M13) methods are approximately similar. Only when combined with RFR the WIS values for the CPS methods are often higher than for the CP methods. CPS provides additional flexibility as a point prediction is not necessarily centered in the middle of an interval. Therefore, it is unexpected that the CP methods that output prediction intervals perform equally well or better than the CPS methods.

During pre-processing, the PV dataset has not been fully subjected to a quality check. However, this dataset has already been pre-processed with the help of a publicly available package [32]. A visual inspection of the hourly aggregated PV data showed no anomalies. Besides, the 175 PV systems have been aggregated which results in increased predictability since extreme values are leveled out and thus affect the quantified uncertainty. Prediction is expected to become more difficult, and thus more uncertain, when using single PV systems or smaller scale aggregated PV systems. This would increase the added value of using probabilistic forecasting compared to point predictions.

For the weighted CP method, the window of the k points to be used for calibration is shifted by 24 timestamps, because bids in the day-ahead market are placed roughly 24 h before delivery. In practice, the bids are not placed 24 h before delivery, but 12 till 36 h before delivery, seeing that all bids are placed at noon. For higher values of K, the impact of this simplification will be negligible, but for smaller values of k this can have an impact on the results. For points just after noon, there is less information for calibration in reality than is accounted for in the simulation. While for points right before noon there is more information in reality. This means that, due to this assumption, the hours just before noon are not estimated as accurately as possible and have wider confidence intervals than they would have had with a dynamic shift of 12 to 36 timestamps while for hours just after noon, the opposite holds.

4. Conclusions and future work

This study contributes to the existing knowledge on solar PV power forecasting by exploring the potential of a forecasting framework based on CP as a novel probabilistic forecasting method to enhance the reliability of day-ahead PV power predictions. Three point prediction models, namely SLR, MLR and RFR, were employed to generate day-ahead point predictions of PV power. Subsequently, various LQR and CP methods are used to quantify the uncertainty of the point predictions in the form of prediction intervals or CDFs. One of the main conclusions drawn from this paper is that employing CP with KNN and/or binning after a LR model yields superior performance compared to the corresponding LQR model. The most accurate representation of uncertainty in this study was obtained when employing RFR in combination with CP using KNN with fifty nearest neighbors and fifteen Mondrian bins. This method led to a 14.0% improvement in the weighted interval score compared to MLR. These findings underscore the potential of utilizing CP to quantify the uncertainty of day-ahead PV power predictions.

With the increasing amount of publications on CP, it is anticipated that CP will continue to be explored and developed in the coming years, further enhancing its potential. Future research can build upon this study by exploring additional variants of CP, particularly additional variants of CQR. To further improve the quantification of uncertainty, it is of added value to use a predictor dataset with parameters indicating the accuracy of the prediction which could be provided by ECMWF. Such accuracy indicators can be used as uncertainty scalars for CP.

In a more practical sense, the proposed CP-based framework leads to more reliable information about the probability distribution of PV power and thus enhanced power predictions. This information can be leveraged in market bidding strategies to increase financial gain for PV power suppliers while mitigating associated risk. Additionally, grid operators can use the uncertainty quantification methods, in combination with the point prediction models, to gain insight in the expected grid imbalance in transmission networks or congestion in distribution networks.

CRediT authorship contribution statement

Yvet Renkema: Conceptualization, Data curation, Investigation, Methodology, Validation, Visualization, Writing – original draft. **Lennard Visser:** Data curation, Validation, Writing – review & editing. **Tarek AlSkaif:** Conceptualization, Formal analysis, Funding acquisition, Project administration, Resources, Supervision, Validation, Writing – review & editing.

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.

References

[1] IEA, World Energy Outlook 2020, Paris, 2020, https://www.iea.org/reports/world-energy-outlook-2020.

[2] B.N. Stram, Key challenges to expanding renewable energy, Energy Policy 96 (2016) 728–734, http://dx.doi.org/10.1016/j.enpol.2016.05.034, URL https://www.sciencedirect.com/science/article/pii/S0301421516302646.

[3] B. Li, J. Zhang, A review on the integration of probabilistic solar forecasting in power systems, Solar Energy 210 (2020) 68–86.

[4] D. Birkeland, T. AlSkaif, Research areas and methods of interest in European intraday electricity market research—A systematic literature review, Sustain. Energy Grids Netw. (2024) 101368.

[5] C. Voyant, G. Notton, S. Kalogirou, M.-L. Nivet, C. Paoli, F. Motte, A. Fouilloy, Machine learning methods for solar radiation forecasting: A review, Renew. Energy 105 (2017) 569–582.

[6] M.N. Akhter, S. Mekhilef, H. Mokhlis, N. Mohamed Shah, Review on forecasting of photovoltaic power generation based on machine learning and metaheuristic techniques, IET Renew. Power Gener. 13 (7) (2019) 1009–1023, http://dx.doi.org/10.1049/iet-rpg.2018.5649.

[7] D. Yang, J. Kleissl, C.A. Gueymard, H.T. Pedro, C.F. Coimbra, History and trends in solar irradiance and PV power forecasting: A preliminary assessment and review using text mining, Sol. Energy 168 (2018) 60–101, http://dx.doi.org/10.1016/j.solener.2017.11.023, Advances in Solar Resource Assessment and Forecasting. URL https://www.sciencedirect.com/science/article/pii/S0038092X17310022.

[8] L. Visser, T. AlSkaif, W. van Sark, Operational day-ahead solar power forecasting for aggregated PV systems with a varying spatial distribution, Renew. Energy 183 (2022) 267–282, http://dx.doi.org/10.1016/j.renene.2021.10.102.

[9] D.W. Van der Meer, J. Widén, J. Munkhammar, Review on probabilistic forecasting of photovoltaic power production and electricity consumption, Renew. Sustain. Energy Rev. 81 (2018) 1484–1512.

[10] L. Massidda, F. Bettio, M. Marrocu, Probabilistic day-ahead prediction of PV generation. A comparative analysis of forecasting methodologies and of the factors influencing accuracy, Sol. Energy 271 (2024) 112422.

[11] A.N. Angelopoulos, S. Bates, A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2021, http://dx.doi.org/10.48550/ARXIV.2107.07511, arXiv URL https://arxiv.org/abs/2107.07511.

[12] C. Kath, F. Ziel, Conformal prediction interval estimation and applications to day-ahead and intraday power markets, Int. J. Forecast. 37 (2) (2021) 777–799, http://dx.doi.org/10.1016/j.ijforecast.2020.09.006, URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85095765900&doi=10.1016%2fj.ijforecast.2020.09.006&partnerID=40&md5=2d087a00b9a00be9771c138d0b07bf5e.

[13] V. Jensen, F.M. Bianchi, S.N. Anfinsen, Ensemble conformalized quantile regression for probabilistic time series forecasting, IEEE Trans. Neural Netw. Learn. Syst. (2022) 1–12, http://dx.doi.org/10.1109/TNNLS.2022.3217694, URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85141644614&doi=10.1109%2fTNNLS.2022.3217694&partnerID=40&md5=6e90ba6f73b3dedc94963c7366943d6f.

[14] S. Tajmouati, B. E.L. Wahbi, M. Dakkon, Applying regression conformal prediction with nearest neighbors to time series data, Comm. Statist. Simulation Comput. (2022) http://dx.doi.org/10.1080/03610918.2022.2057538, URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85129220059&doi=10.1080%2f03610918.2022.2057538&partnerID=40&md5=4083e8e46c9f99037ed3a81d79a0efcd.

[15] J. Hu, Q. Luo, J. Tang, J. Heng, Y. Deng, Conformalized temporal convolutional quantile regression networks for wind power interval forecasting, Energy 248 (2022) 123497, http://dx.doi.org/10.1016/j.energy.2022.123497, URL https://www.sciencedirect.com/science/article/pii/S0360544222004005.

[16] W. Wang, B. Feng, G. Huang, C. Guo, W. Liao, Z. Chen, Conformal asymmetric multi-quantile generative transformer for day-ahead wind power interval prediction, Appl. Energy 333 (2023) 120634, http://dx.doi.org/10.1016/j.apenergy.2022.120634, URL https://www.sciencedirect.com/science/article/pii/S0306261922018918.

[17] L. Massidda, M. Marrocu, Total and thermal load forecasting in residential communities through probabilistic methods and causal machine learning, Appl. Energy 351 (2023) 121783, http://dx.doi.org/10.1016/j.apenergy.2023.121783, URL https://www.sciencedirect.com/science/article/pii/S0306261923011479.

[18] M. Jain, T. AlSkaif, S. Dev, Are deep learning models more effective against traditional models for load demand forecasting? in: 2022 International Conference on Smart Energy Systems and Technologies, SEST, 2022, pp. 1–6, http://dx.doi.org/10.1109/SEST53650.2022.9898424.

[19] D.V. Pombo, H.W. Bindner, S.V. Spataru, P.E. Sørensen, P. Bacher, Increasing the accuracy of hourly multi-output solar power forecast with physics-informed machine learning, Sensors 22 (3) (2022) http://dx.doi.org/10.3390/s22030749, URL https://www.mdpi.com/1424-8220/22/3/749.

[20] L. Grinsztajn, E. Oyallon, G. Varoquaux, Why do tree-based models still outperform deep learning on tabular data?, 2022, http://dx.doi.org/10.48550/ARXIV.2207.08815, arXiv URL https://arxiv.org/abs/2207.08815.

[21] S. Elsayed, D. Thyssens, A. Rashed, H.S. Jomaa, L. Schmidt-Thieme, Do we really need deep learning models for time series forecasting?, 2021, http://dx.doi.org/10.48550/ARXIV.2101.02118, arXiv URL https://arxiv.org/abs/2101.02118.

[22] Y. Grushka-Cockayne, V. Jose, The M4 forecasting competition prediction intervals, SSRN Electron. J. (2019) http://dx.doi.org/10.2139/ssrn.3329413.

[23] T. AlSkaif, S. Dev, L. Visser, M. Hossari, W. van Sark, A systematic analysis of meteorological variables for PV output power estimation, Renew. Energy 153 (2020) 12–22, http://dx.doi.org/10.1016/j.renene.2020.01.150, URL https://www.sciencedirect.com/science/article/pii/S0960148120301725.

[24] 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, http://dx.doi.org/10.1016/j.apenergy.2022.119598, URL https://www.sciencedirect.com/science/article/pii/S0306261922009059.

[25] D. Markovics, M.J. Mayer, Comparison of machine learning methods for photovoltaic power forecasting based on numerical weather prediction, Renew. Sustain. Energy Rev. 161 (2022) 112364, http://dx.doi.org/10.1016/j.rser.2022.112364, URL https://www.sciencedirect.com/science/article/pii/S136403212200274X.

[26] U. Johansson, H. Boström, T. Löfström, H. Linusson, Regression conformal prediction with random forests, Mach. Learn. 97 (1–2) (2014) 155–176, http://dx.doi.org/10.1007/s10994-014-5453-0, URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-84906946396&doi=10.1007%2fs10994-014-5453-0&partnerID=40&md5=39f3763fee5c90c135a3f93e730e2b99 All Open Access, Bronze Open Access, Green Open Access.

[27] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.

[28] E. Morgan, R. Zhou, W. Feng, Prediction intervals of machine learning models for taxi trip length, 343, 2021, pp. 715–724, http://dx.doi.org/10.1007/978-3-030-63591-6_65, URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85115163703&doi=10.1007%2f978-3-030-63591-6_65&partnerID=40&md5=de5a69f2b2d83475bb292bada56a0eb9,

[29] H. Boström, Conformal regressors and predictive systems – a gentle introduction, 2022, p. 7, URL https://cml.rhul.ac.uk/copa2022/presentations/COPA_2022_Presentation_Conformal_Regressors_and_Predictive_Systems__a_Gentle_Introduction.pdf.

[30] H. Boström, U. Johansson, T. Löfström, Mondrian conformal predictive distributions, Conformal Probabilistic Predict. Appl. 152 (2021) 1–15, URL https://proceedings.mlr.press/v152/bostrom21a/bostrom21a.pdf.

[31] H. Boström, Conformal regressors and predictive systems – a gentle introduction, 2022, pp. 20–28, URL https://cml.rhul.ac.uk/copa2022/presentations/COPA_2022_Presentation_Conformal_Regressors_and_Predictive_Systems__a_Gentle_Introduction.pdf.

[32] L.R. Visser, B. Elsinga, T.A. AlSkaif, W.G.J.H.M. van Sark, Open-source quality control routine and multi-year power generation data of 175 PV systems, J. Renew. Sustain. Energy 14 (4) (2022) 043501, http://dx.doi.org/10.1063/5.0100939.

[33] G. ECMWF, European centre for medium-range weather forecasts, ECMWF, 2020.

[34] 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, Sol. Energy 251 (2023) 86–105, http://dx.doi.org/10.1016/j.solener.2023.01.019, URL https://www.sciencedirect.com/science/article/pii/S0038092X23000191.

[35] W.F. Holmgren, C.W. Hansen, M.A. Mikofski, pvlib python: A python package for modeling solar energy systems, J. Open Source Softw. 3 (29) (2018) 884.

[36] J. Bracher, E. Ray, T. Gneiting, N. Reich, Evaluating epidemic forecasts in an interval format, PLoS Comput. Biol. 17 (2021) e1008618, http://dx.doi.org/10.1371/journal.pcbi.1008618.