Solar Energy 251 (2023) 86–105

Contents lists available at ScienceDirect

Solar Energy

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

On the value of expert knowledge in estimation and forecasting of solar photovoltaic power generation

Lennard Visser [a,*], Tarek AlSkaif [c], Jing Hu [a], Atse Louwen [b], Wilfried van Sark [a]

[a] Copernicus Institute of Sustainable Development, Utrecht University, Princetonlaan 8a, 3584 CB Utrecht, The Netherlands
[b] Institute for Renewable Energy, EURAC Research, Viale Druso 1, 39100, Bolzano, Italy
[c] Information Technology Group, Wageningen University and Research, Droevendaalsesteeg 4, 6708 PB Wageningen, The Netherlands

ARTICLE INFO

Keywords:
Photovoltaics
Solar power forecasting
Solar power estimation
Meteorological variables
Predictor variables

ABSTRACT

Reliable estimates and forecasts of Photovoltaic (PV) power output form a fundamental basis to support its large-scale integration. This is recognized in literature, where a growing amount of studies deal with the development of PV power estimation and forecasting models. In particular, machine learning techniques received significant attention in the past decade. Yet, the importance of predictor variables are consistently ignored in such developments and as a result those models fail to acknowledge the value of including physics-based models. In this study we quantify the value of predictor variables for PV power estimation and forecasting, assess deficiencies in estimation and forecasting models, and introduce a number of pre-processing steps to improve the overall estimation or forecasting performance. To this end, we use common physical models to create so-called expert variables and test their impact on the performance of single-point and probabilistic models. In addition, we investigate the optimal selection of predictor variables for PV power estimation and forecasting. By means of a sensitivity analysis, the paper shows how the value of expert variables is affected by the tilt angle of the PV system. To allow for a deeper insight into the importance of predictor variables, two case studies in different climate regions are considered in the numerical evaluation.

1. Introduction

Accurate estimates and forecasts of potential power production of Photovoltaic (PV) systems are essential to host their rapidly growing capacity in the electricity grid (IEA, 2020). Solar power estimates are needed to foresee the potential contribution of new PV systems to the (local) power supply, and calculate its impact on the electricity grid. Forecasts can improve the dispatch of electricity generation, and subsequently limit the reserve capacity needed to maintain grid stability. Hence, reliable PV power estimation and forecasting models play a vital role in cost-effective grid operation (Visser et al., 2022e). There is a growing body of literature that recognizes the importance of such models for effective large-scale integration of PV systems into the grid. Several methods for solar power estimation and forecasting have been developed so far. State-of-the-art solar estimation models typically rely on weather measurements and/or reanalysis data. The main source of information of solar forecasting models depends on the time horizon of interest, and typically consider weather data from all-sky imaging, satellite imaging and/or Numerical Weather Predictions (NWP). The most successful models post-process such data using a wide variety of models (e.g. physical, regression/statistical and machine learning models), possibly combining them together resulting in ensemble or hybrid forecast approaches (Nguyen and Müsgens, 2022). The preferred approach, i.e. the optimal combination of data source and model, is highly dependent on the forecast horizon of interest (Yang et al., 2022).

Three main directions in the field of solar forecasting can be identified, namely: (i) advanced forecast models, which focus on machine learning, extensive data collection and/or data manipulation techniques (Rana and Rahman, 2020), (ii) probabilistic models that describe the forecast uncertainty (Van der Meer et al., 2018) and (iii) relative new approaches like firm power forecasts, which aim to eliminate the forecast uncertainty by considering a PV–battery system (Perez et al., 2020). These study directions share a common foundation, which is to establish a relation between a (number of) predictor variable(s) and the target variable, i.e. the PV power output. Despite their pivotal role, to the authors' best knowledge, hardly any studies have dealt with the relative contribution or importance of (specific) predictor variables (Yang and van der Meer, 2021). In related fields to solar

* Corresponding author.
E-mail addresses: l.r.visser@uu.nl (L. Visser), tarek.alskaif@wur.nl (T. AlSkaif), J.Hu@uu.nl (J. Hu), atse.louwen@eurac.edu (A. Louwen), W.G.J.H.M.vanSark@uu.nl (W. van Sark).

https://doi.org/10.1016/j.solener.2023.01.019
Received 5 September 2022; Received in revised form 9 December 2022; Accepted 11 January 2023
Available online 18 January 2023

forecasting, such as wind and electricity price forecasting, predictor variables receive more attention. For example, Díaz et al. (2019) and Visser et al. (2020) performed a comprehensive study into the importance of a set of predictor variables related to the formation of electricity spot market prices in Spain and the Netherlands, respectively. Alternatively, power curves that establish a relation between a predictor and target variable (i.e. wind speed and wind power), gained significant attention in the field of wind forecasting (Jeon and Taylor, 2012; Xu et al., 2016; Yang and van der Meer, 2021). Similar studies are absent in the present literature on solar estimation and forecasting (Yang and van der Meer, 2021). An exemption is found in AlSkaif et al. (2020), where the interdependence of a number of variables and their importance for solar power estimation is studied. Yet, this study does not cover the contribution of amongst others global horizontal irradiance (G_{GHI}), which may be the single most used variable in solar estimation and forecasting. Other commonly consulted variables include temperature, cloud cover, precipitation and wind speed (Singla et al., 2021; Ahmed et al., 2020; Sobri et al., 2018).

Another aspect that is commonly overlooked in the development of (regression/statistical and machine learning based) solar estimation and forecasting models is the creation of additional variables that carry potential value to the model, but are commonly not readily available. The power output of a PV system is directly dependent on the direct normal irradiance (G_{DNI}) and diffuse horizontal irradiance (G_{DHI}) received on the plane of the PV array. Nevertheless, measurements at weather stations and weather forecasts are usually limited to the G_{GHI} received per square meter of surface. The in-plane irradiance variables can easily be created, as a result of the efforts made by the PV community in the past decades. A number of models were developed that have proven to be very effective in: decomposing the G_{GHI} into G_{DNI} and G_{DHI}, transposing these into irradiance that is received at the plane of the array and converting this information into the expected PV power output (Mayer and Gróf, 2021). We refer to the variables outputted by these models as expert variables. Solar engineers have made many of these decomposition (also referred to as separation), transposition and PV models available in an extensive open-source library called *pvlib* (Holmgren et al., 2018).

Apart from a more accurate performance in terms of classic error metrics, it is expected that including these expert variables in the set-up of estimation and forecasting models improves the models' explainability and predictability. This is because these expert variables improve the representation of the underlying physical and statistical dynamics in PV power generation. In particular, they enable the model to better capture the seasonality, which is a non-stationary time-dependent variation in the PV power output. There are three conventional approaches to account for seasonality in PV power output simulations i.e. through introducing seasonal dummy variables, by considering harmonic regression models or by using the clearness index (Boland, 2020; Lauret et al., 2015; Akhter et al., 2019; Nguyen and Müsgens, 2022). The first two approaches directly introduce seasonal components into the estimation or forecasting model, but they also treat seasonality in a deterministic manner (Young et al., 1999). Harmonic regression with e.g. Fourier terms is particularly good at capturing long-term seasonal cycles, but it also incurs greater mathematical complexity (Hyndman and Athanasopoulos, 2018). Alternatively, seasonality can be removed by replacing the irradiance variables by corresponding indices, which describe the seasonality. For example, by replacing the G_{GHI} with the clearness index, being the ratio between G_{GHI} and clear sky irradiance. After the estimation or forecasting process, the clearness index can be converted back into PV power outputs. However, this approach is particularly susceptible to data input errors, because close-to-zero clear sky irradiance values at and around sunrise and sunset can artificially inflate the calculated clearness index and is in these cases thus very sensitive to small deviations of input values. Compared with these conventional approaches, the direct employment of expert variables in the forecasting and estimation models is suggested to avoid these disadvantages. With this approach, the aim is to describe the seasonality in the model rather than removing it.

As mentioned above, very few studies are found to leverage the value of expert variables for PV power estimation and forecasting, thus ignoring the potential effect on the model performance. Examples of studies that leverage expert variables include (Pombo et al., 2022; Visser et al., 2022a). Both studies evaluated the optimal combination of features with the purpose of PV power forecasting. Nevertheless, the number of features considered is limited and NWP forecasts are, for instance, not included in Pombo et al. (2022). Similarly, Markovics and Mayer (2022) studied the effect of different predictor variable selections, but without consulting physical models to generate the expert variables. Besides, similar to Pombo et al. (2022), Visser et al. (2022a), the total number of predictor variables considered is limited. In contrast, Mayer (2022) presented a very comprehensive analysis of the contribution of adopting decomposition, transposition and PV models as pre-processing steps by feeding the expert variables to an artificial neural network model. In the study, Mayer (2022) evaluated the forecast quality improvement for a single-point forecast model by incorporating these expert variables.

This paper aims to build upon the work presented in Mayer (2022), by assessing the impact of including expert variables in both probabilistic and single-point, estimation and forecasting models. Similar to Mayer (2022), specific attention is given to the contribution of expert variables, i.e. variables outputted by the decomposition, transposition and PV models, to the performance of the forecasting and estimation models. In this context, linear regression and nonlinear machine learning models are tested and compared to two benchmark models. While the work of Mayer (2022) is limited to evaluating the performance improvement by including such expert variables in a single-point forecast model, the work in this paper considers single-point and probabilistic, estimation and forecasting models. Next, this study exposes the underlying model dynamics that explain the improvement. In addition, we evaluate the optimal selection of predictor variables for the purpose of PV power estimation and forecasting. Since the contribution of including a transposition model is in particular dependent on the tilt angle of the PV system, a sensitivity analysis of the model improvement to expert variables as a function of the tilt angle is conducted. In order to enhance the robustness and interpretability of the paper results, two different case studies with different climates are considered. Consequently, all experiments are simultaneously conducted for Utrecht, the Netherlands, and Bolzano, Italy, representing a temperate oceanic and a humid subtropical climate, respectively.

The structure of the paper is organized as follows. Section 2 presents the research methods. The data is described in Section 3. Section 4 presents and discusses the main findings of this research. Lastly, the paper is concluded in Section 5.

2. Methods

2.1. Data manipulation

The collected data per case study span a period of three years and are represented as a matrix **X**. All night time values are discarded from **X** by filtering for positive clear sky irradiance values. Next, to improve the model accuracy and reduce computational demands, all predictor variables are normalized to values between 0 and 1. The target variable, i.e. the PV power output (**p**), is normalized to the reported installed AC capacity. Then, **X** is split into training and test sets, where the training set holds all values in the first two years of the dataset and the test set contains the last year of the dataset. In this research, all models are fitted to the training data and evaluated using the test data. Where relevant, the training data is further split into training and validation sets for the purpose of hyperparameter tuning, using k-fold cross-validation ($k = 8$) (Raschka and Mirjalili, 2019). The data collection for both case studies is discussed in Section 3.

2.2. Estimation and forecasting models

2.2.1. Single-point models

To explore the value of (expert) variables in solar estimation and forecasting, linear, nonlinear and benchmark models are considered to produce both single-point and probabilistic estimates and forecasts of the PV power output given a variety of predictor variables. Single-point models are also referred to as single-valued models, i.e. models that output a single value. The first model that is considered in this study is a Multivariate Linear Regression (MLR) model. MLR is a simple yet effective model that is widely adopted to forecast and estimate the PV power output (AlSkaif et al., 2020; Visser et al., 2022b; Markovics and Mayer, 2022). 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, i.e. the PV power output. In this study, the least squares loss function is used.

Secondly, we include a Random Forest regression (RF) model, which is a nonlinear model that has proven its value in the field of solar estimation and forecasting in previous studies (AlSkaif et al., 2020; Visser et al., 2022b; Pombo et al., 2022). RF is an ensemble based model that consists of a number of trees, each made up of t layers and 2^t decision nodes. 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 variables is considered to construct the decision nodes by optimizing on a loss function, e.g. least squares (Breiman, 2001). The output of an RF model is equal to the conditional mean of all constructed trees.

The performance of the single-point forecast models mentioned above is benchmarked using a clear sky persistence (CSP) and a physical PV model (PV_{lib}). The output of the CSP model is set equal to the most recent observation of the same time in a previous day, corrected by the change in clear sky irradiance. Since the estimation models consider current observations instead of predictions, this benchmark model would give perfect estimates of the PV power generation and is therefore not included in the evaluation of the solar estimation models.

The use of a PV model to obtain the PV power output involves few sub-steps (Visser et al., 2022e). First, the Erbs decomposition model is employed to extract the G_{DNI} and G_{DHI} from the G_{GHI} (Erbs et al., 1982). Second, with the Perez transposition model we obtain respectively the global, direct, diffuse, ground diffuse and sky diffuse irradiance received on the plane of the PV array ($G_{A.GI}$, $G_{A.BI}$, $G_{A.DI}$, $G_{A.DI_G}$ and $G_{A.GI_S}$), which requires G_{DNI}, G_{DHI} and G_{GHI} as input (Perez et al., 1990). Finally, the CEC PV (Holmgren et al., 2018) and Sandia inverter (Boyson et al., 2007) models are used to subsequently model the DC ($P_{PV.DC}$) and AC (P_{PV}) power output of a PV system based on the system's characteristics and relevant weather variables, including the irradiance components, temperature and wind speed (Holmgren et al., 2018). In addition, the Sandia model is used to generate the operating cell temperature of the PV module(s) (T_{PV}) (Kratochvil et al., 2004). An overview of all the variables and the models used to create these is given in Table 2.

2.2.2. Probabilistic models

The first probabilistic model we consider in this study is quantile regression (QR). Similar to the MLR model, QR establishes a linear relationship between the predictor and target variables. In QR, the coefficients are learned independently per percentile τ by minimizing the sum of absolute residuals over the asymmetrically applied weights error (Koenker and Hallock, 2001). QR is widely applied and proved its value in the field of solar forecasting (Lauret et al., 2017; Brinkel et al., 2021). In this study we consider a 99% prediction interval using a total of 21 quantiles, these describe an interval of 5% from 5 to 95% and include a lower and upper bound of 0.5 to 99.5%.

Secondly, we consider a quantile regression forest (QRF) model. The QRF model operates similarly to the RF model, where the mean value per node is replaced by the distribution of observations. As a result, the QRF model gives the conditional distribution function or weighted distribution of observations (Meinshausen and Ridgeway, 2006). The value of adopting a QRF model for the purpose of PV power forecasting is shown by Tripathy et al. (2020).

Finally, the clear sky persistence ensemble (CSPE) model is used to benchmark the results obtained by the probabilistic models (Pedro et al., 2018). The output of the CSPE model is obtained similar to the CSP model, where instead of a single value the 21 most recent observations are considered that together form a distribution.

Table 1
Overview of the predictor variables considered per estimation and forecasting model configuration. Note that the written abbreviations can be found in Table 2.

Model configuration	Variable category	Variables included
1	Standard variables	AM_A, AM_{MSL}, CC_T, G_{GHI}, WS_{u10}, WS_{v10}, T_A, T_D, TP
2	Decomposition variables	**1** & G_{DNI}, G_{DHI}, AM_a, AM_r
3	Transposition variables	**2** & $G_{A.BI}$, $G_{A.DI}$, $G_{A.DI_G}$, $G_{A.DI_S}$, $G_{A.GI}$
4	PV model variables	**3** & T_{PV}, P_{PV}

2.2.3. Model input

The value of expert variables (which are discussed in more detail in Section 3.2) in this study is primarily quantified by considering the performance improvement of the estimation and forecasting models using the error metrics presented in Section 2.3. The value is assessed by introducing the expert variables in the models step by step, denoting these as model configurations **1** to **4** respectively. First, a reference model configuration **1** is tested, which relies on standard weather variables only. Next, the outputted variables of the decomposition model are added, resulting in model configuration **2**. In model configurations **3** the variables generated by the transposition model are incorporated. Model configuration **4** incorporates the PV power output and cell temperature, which are produced using the CEC and Sandia models. These steps are summarized in Table 1. Note, the number of predictor variables are reduced for the QR_3 and QR_4 models, see Section 3.3.

2.3. Error metrics

2.3.1. Single-point models

The models that produce single-point estimates and forecasts are evaluated on the mean absolute error (MAE), root mean square error (RMSE) and bias, see Eq. (1), (2) and (3). These metrics are considered as they are commonly used for evaluating single-point forecasting and estimation models (Ahmed et al., 2020).

$$MAE = \frac{1}{T}\sum_{t=1}^{T} y_t - \hat{y}_t, \tag{1}$$

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}, \tag{2}$$

$$Bias = \frac{1}{T}\sum_{t=1}^{T} y_t - \hat{y}_t, \tag{3}$$

where y and $\hat{y}$ present the observations and estimated or forecasted normalized PV power output per time-step t in T.

2.3.2. Probabilistic models

The statistical metrics presented above can also be adopted for evaluating probabilistic estimates and forecasts if a single value, e.g. mean or median, is extracted. However, to evaluate the model output on their ability to quantify the uncertainty, other metrics must be adopted. The performance of probabilistic models is typically characterized by its

Table 2
Overview of the collected and constructed variables. The collected variables present weather variables that are directly retrieved from its source, e.g. weather forecasts. Constructed variables are processed variables, derived through the use of physical models.

	Abbreviation	Variable	Unit	Type	Method	Reference
1.	AM_A	Absolute air mass	–	Constructed	$AM_R \frac{P_A}{101325}$	Holmgren et al. (2018)
2.	AM_R	Relative air mass	–	Constructed	Kastenyoung model	Kasten and Young (1989)
3.	CC_T	Total cloud cover	[0-1]	Collected		ECMWF (2020) (Muñoz Sabater, 2019)
4.	G_{CSI}	Clear sky irradiance	W/m^2	Constructed	Ineichen model	Ineichen and Perez (2002)
5.	G_{DNI}	Direct normal irradiance	W/m^2	Constructed	Erbs model	Erbs et al. (1982)
6.	G_{DHI}	Diffuse horizontal irradiance	W/m^2	Constructed	Erbs model	Erbs et al. (1982)
7.	G_{ETR}	Extraterrestrial irradiance	W/m^2	Constructed	Spencer method	Spencer (1982)
8.	G_{GHI}	Global horizontal irradiance	W/m^2	Collected		ECMWF (2020) (KNMI, 2020)
9.	$G_{A.GI}$	Total in-plane irradiance (i.e. Global tilted irradiance)	W/m^2	Constructed	Perez model	Perez et al. (1990)
10.	$G_{A.BI}$	Total in-plane beam irradiance	W/m^2	Constructed	Perez model	Perez et al. (1990)
11.	$G_{A.DI}$	Total in-plane diffuse irradiance	W/m^2	Constructed	Perez model	Perez et al. (1990)
12.	$G_{A.DI_S}$	In-plane diffuse irradiance from sky	W/m^2	Constructed	Perez model	Perez et al. (1990)
13.	$G_{A.DI_G}$	In-plane diffuse irradiance from ground	W/m^2	Constructed	Perez model	Perez et al. (1990)
14.	P_{PV}	Inverter model estimated AC power output of a PV system	W	Constructed	Sandia model	Boyson et al. (2007)
15.	$P_{PV.DC}$	PV model estimated DC power output of a PV system	W	Constructed	CEC model	Holmgren et al. (2018)
16.	P_A	Atmospheric pressure	Pa	Collected		ECMWF (2020) (Muñoz Sabater, 2019)
17.	P_{MSL}	Mean sea level pressure	Pa	Collected		ECMWF (2020) (Muñoz Sabater, 2019)
18.	T_A	Ambient temperature	°C	Collected		ECMWF (2020) (Muñoz Sabater, 2019)
19.	T_D	Dewpoint temperature	°C	Collected		ECMWF (2020) (Muñoz Sabater, 2019)
20.	T_{PV}	Cell temperature	°C	Constructed	Sandia model	Kratochvil et al. (2004)
21.	TP	Total precipitation	m	Collected		ECMWF (2020) (Muñoz Sabater, 2019)
22.	WS_{u10}	Zonal wind speed at 10 m height	m/s	Collected		ECMWF (2020) (Muñoz Sabater, 2019)
23.	WS_{v10}	Meridional wind speed at 10 m height	m/s	Collected		ECMWF (2020) (Muñoz Sabater, 2019)
24.	WS	Wind speed	m/s	Constructed	$\sqrt{WS_{u10}{}^2 + WS_{v10}{}^2}$	
25.	θ_z	Solar zenith angle	°	Constructed	NREL solar position method	Reda and Andreas (2004)
26.	σ_A	Solar azimuth angle	°	Constructed	NREL solar position method	Reda and Andreas (2004)

Table 3
The top five most important predictor variables for the single-point estimation and forecasting models at de Bilt and Bolzano. The most important variable is ranked No. 1. The values overlap with the order of included predictor variables as depicted on the x-axis in Fig. 7. A complete overview is given in Table B.1 in Appendix B.

No.	Estimation				Forecasting			
	De Bilt, NL		Bolzano, IT		De Bilt, NL		Bolzano, IT	
	MLR	RF	MLR	RF	MLR	RF	MLR	RF
#1	P_{PV}	P_{PV}	P_{PV}	P_{PV}	P_{PV}	P_{PV}	P_{PV}	P_{PV}
#2	G_{GHI}	$G_{A.DI_G}$	G_{GHI}	G_{GHI}	WS_{u10}	$G_{A.DI_G}$	G_{GHI}	AM_R
#3	G_{DNI}	G_{DNI}	$G_{A.GI}$	G_{DNI}	G_{DNI}	G_{DNI}	G_{DNI}	CC_T
#4	$G_{A.GI}$	P_{MSL}	T_D	T_D	$G_{A.GI}$	CC_T	CC_T	P_{MSL}
#5	WS_{v10}	AM_A	G_{DHI}	CC_T	AM_A	TP	AM_A	$G_{A.DI_S}$

reliability and sharpness (Van der Meer et al., 2018). The reliability of the model is assessed in this study using the prediction interval coverage probability (PICP), which indicates the rate at which the forecast covers the observations, see Eq. (4). A high PICP indicates that more values lie within the bounds of the prediction interval. A PICP value that is approximately equal to the prediction interval is preferred.

The PICP metric is complemented with the prediction interval normalized average width (PINAW), see Eq. (5). In contrast to the PICP,

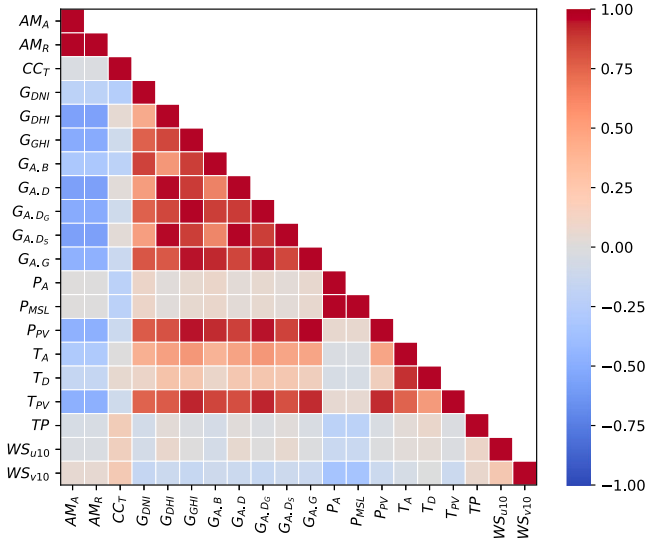


Fig. 1. Cross-correlation values for the forecasted predictor variables at de Bilt, considering three years of data.

PINAW evaluates the width, i.e. the range between the lower and upper bound, of the model output. Consequently, PINAW assesses the width of the prediction intervals.

Unlike the PICP and PINAW, the continuous ranked probability score (CRPS) captures both the reliability and sharpness of a probabilistic model, see Eq. (6). $F_t(x)$ and $\hat{F}_t(x)$ are the cumulative distribution functions of the observations and estimates or forecasts of the PV power output at time-step t in T. Note that $F_t(x)$ is a cumulative-probability step function as it describes a single-value, which jumps from 0 to 1 where the forecast variable x is equal to the observation (Lauret et al., 2019). Particularly, CRPS rewards a high concentration of the estimated or forecasted probability around the target value. Therefore, the CRPS may be used as a global metric, where a lower CRPS value is an indicator for a more accurate model. Besides, for a single-point model, the CRPS reduces to the MAE (Yang and van der Meer, 2021). The CRPS is therefore comparable to the MAE.

$$PICP = \frac{1}{T}\sum_{t=1}^{T} \epsilon, \longmapsto \epsilon = \left\{ \begin{array}{l} 1 \text{ if } \hat{y}_t \in [L_t, U_t] \\ 0 \text{ if } \hat{y}_t \notin [L_t, U_t] \end{array} \right\}, \tag{4}$$

$$PINAW = \frac{1}{TR}\sum_{t=1}^{T}(U_t - L_t), \tag{5}$$

$$CRPS = \frac{1}{T}\sum_{t=1}^{T}\int_{-\infty}^{\infty}(F_t(x) - \hat{F}_t(x))^2 dx, \tag{6}$$

where R is equal to the difference between the maximum and minimum estimated or forecasted value. U and L present the upper and lower bound of the prediction interval.

2.4. Feature selection

The optimal selection of features for solar estimation and forecasting is obtained by means of forward feature selection (Bemister-Buffington et al., 2020). Forward feature selection comprises an iterative process where the model is firstly fit with each single feature, separately. The feature, i.e. predictor variable, that supports the best performing model is kept in a selected feature list and considered the most valuable feature. As a next step the model is fit with two features, i.e. a combination of the most valuable feature listed in the selected feature list and each remaining feature. The second feature of the best performing model is identified and added to the list of selected features. This process is repeated until all features are included in the model.

The forward feature selection is performed separately for the MLR, RF, QR and QRF model, for the purpose of solar power estimation and forecasting in both case studies. The selection is based on the (lowest) obtained MAE or CRPS score (see Section 2.3).

2.5. Sensitivity analysis

In the sensitivity analysis we evaluate how the value of considering expert variables in the estimation and forecasting models is affected by the tilt angle of the PV system. This analysis is conducted for both case studies, considering a tilt angle interval of 10° ranging from 0° to 90°. Instead of actual PV power measurements, this analysis relies on the PV power output estimates generated using the PV model introduced in Section 2.2.1, with the reason being the unavailability of PV power measurements at the same site with varying tilt angles. This means that for the sensitivity analysis, y is, for both estimation and forecasting, set equal to P_{PV} calculated using weather measurements.

3. Data collection and analysis

3.1. PV power output

For Utrecht, the PV power output is collected from a 2.7 kWp DC, 2.6 kWp AC PV system (data from PV system with ID107 in Visser et al. (2022d,c)). The orientation of this PV system is defined by a tilt angle of 32° and an azimuth angle of 180°, the optimal orientation for a PV system in the Netherlands was previously identified with a tilt angle of 37° and an azimuth angle of 180° (Louwen et al., 2017). For Bolzano, the PV power output is collected from a 4.2 kWp DC, 4.0 kWp AC PV system. The orientation of the PV system is defined by a tilt angle of 30° and an azimuth angle of 188.5°, the optimal orientation for a PV system in Northern Italy was previously identified with a tilt angle of 35–40° and an azimuth angle of 180° (Louwen et al., 2017). For both systems, the power values are normalized to the installed AC capacity, to obtain only production values between 0 and 1.

3.2. Predictor variables

An overview of the variables considered in this study is presented in Table 2. The variables can be split in two main categories, namely: (i) variables that are openly available and can be retrieved online in the form of measurements and forecasts; and (ii) variables that are constructed using open access models developed by solar engineers and recorded in the python package *pvlib* (Holmgren et al., 2018). In this study, all the variables are collected four times in total. Namely, for each study location (i.e. de Bilt and Bolzano), both measurements and forecasts of the variables are collected.

The retrieved predictor variables for the PV power forecasting models concern NWP. For both case studies these variables are retrieved from the European Centre for Medium-Range Weather Forecasts (ECMWF) weather archive (ECMWF, 2020). The predictions are generated by the high resolution (HRES) forecast model of the Integrated Forecast System (IFS) developed by ECMWF.

The variables considered in the PV power estimation models are, except for G_{GHI}, collected from the ERA5 database provided by ECMWF (Muñoz Sabater, 2019). For accuracy reasons, we collect G_{GHI} values from local measurements. For Utrecht, hourly G_{GHI} measurements are retrieved from a KNMI (Royal Netherlands Meteorological Institute) weather station in De Bilt (located in the province of Utrecht, **52°10'N, 5°18'E**) (KNMI, 2020), at circa 9 kilometers from the PV system. For Bolzano, G_{GHI} values are obtained with a 15-min temporal resolution from a pyranometer located less than 50 meters away from the PV system.

The expert variables comprise the (intermediate) outputs of the decomposition, transposition and PV model, which were introduced in Section 2.2.1 for the construction of the PV model, i.e. PV_{lib}. Thus,

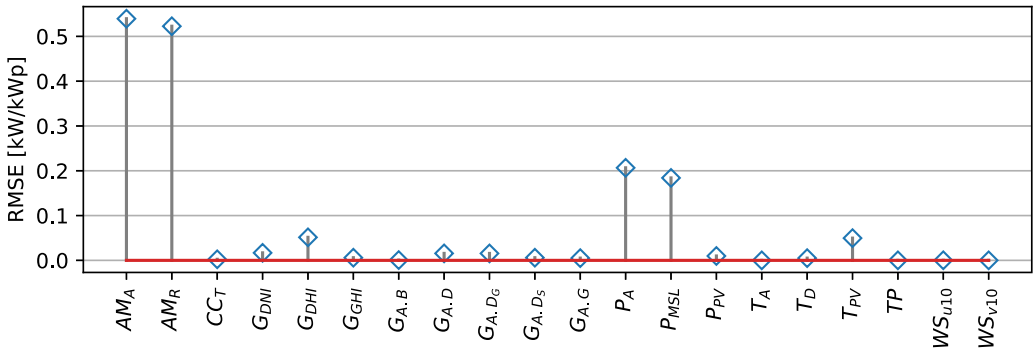


Fig. 2. Predictor variable importance defined as feature permutation importance (Altmann et al., 2010). The y-axis value presents the loss in model performance as a result of randomly shifting the test data of a single predictor variable.

the decomposition variables are the G_{DNI}, G_{DHI}, absolute air mass AM_A and relative air mass AM_R. The transposition variables are the counterparts of the irradiance variables that describe the irradiance received in the plane of the PV array, i.e. total in-plane irradiance $G_{A.GI}$, total in-plane beam irradiance $G_{A.BI}$, total in-plane diffuse irradiance $G_{A.DI}$ and its components: the in-plane diffuse irradiance from the sky $G_{A.DI_S}$ and the in-plane diffuse irradiance from the ground $G_{A.DI_G}$. The PV model variables include the modeled PV power output P_{PV} and the modeled cell temperature of the PV module T_{PV}. An overview of all these constructed variables and the models used to create them is presented in Table 2. In addition, the table presents all other intermediate variables that were needed to create these constructed (expert) variables including the models used to generate them.

3.3. Data analysis

The cross-correlation between predictor variables presents a measure of their linear dependence (AlSkaif et al., 2020). For the development of estimation and forecasting models it is key to be aware of the cross-correlation values amongst the predictor variables as severe multicollinearity between predictor variables may interfere with the model accuracy. Multicollinearity can lead to predictor variables becoming insignificant, whereas their influence on the target variable may be significant. Alternatively, the contribution of two predictor variables can be significant when they balance each other out. Multicollinearity is particularly an issue for models that describe a linear relationship between the predictor and target variables, e.g. MLR and QR. Such cases can easily be identified by observing changes in regression coefficients while changing the selection of predictor variables. To this end, this study investigates the optimal feature selection per model, see Section 2.4.

The cross-correlation values for the entire study period (i.e. three years) is computed between all the predictor variables using Pearson correlation. The correlation values for the forecasted predictor variables at de Bilt are depicted in Fig. 1. The figure presents high positive correlations between the several irradiance variables. Similarly, AM_A and AM_R are highly correlated. Besides, different correlations amongst the temperature variables are noticed, i.e. a significant higher positive correlation between T_A and T_D compared to T_{PV}. The cross-correlations figures for the other estimation and forecasting applications are included in Appendix A. These are accompanied with general statistics of the data.

An example of a multicollinearity issue identified during experiments performed in this study concerns the (high) importance of variables AM_R and AM_A for the MLR forecasting model at de Bilt (see Fig. 2). The importance of these variables is exaggerated as further analysis shows that the variables cancel each other out. This observation is explained by the high positive correlation (~ 1.0) of these variables in combination with large opposing regression coefficients, 3.51 and -3.58 for AM_R and AM_A. This problem was also observed for P_A and P_{MSL}.

The *statsmodels* Python package used to build the QR model does not allow for strong multicollinearity, i.e. near perfect correlation, between the predictor variables. Therefore, in the remainder of this study the least informative predictor variable is dropped when running QR simulations.

4. Results

4.1. Model performance

4.1.1. Single-point models

Fig. 3 shows the performance of the single-point models. The figure presents the results of all single-point models for both applications (i.e. solar PV power output estimation and forecasting) and the two locations (i.e. de Bilt and Bolzano). For both case studies the MAE of all forecast models except CSP lies between 0.06 and 0.10 kW/kWp. The RMSE takes values between 0.10 and 0.13. The forecast models significantly outperform the CSP model, whereas the second benchmark model PV_{lib} performs equivalent. As expected, the estimation models perform significantly better than the forecast models. Here, an MAE and RMSE of 0.01 to 0.07 and 0.03 to 0.10 kW/kWp are found, respectively. Note that no results are displayed for the CSP estimation model (as discussed in Section 2.2.1). Significantly better results are obtained for the estimation models in Bolzano compared to de Bilt. This is explained by a higher temporal resolution of the G_{GHI} measurements at Bolzano and a smaller distance from the location of these measurements to the PV system (see Sections 3.1 and 4.2). Lastly, although the biases are marginal, i.e. for most models between -0.01 and 0.01 kW/kWp, a remarkable pattern is observed for both sites. Except for the PV_{lib} model, a general positive bias is found for all forecast models and a negative bias is observed for all estimation models.

From the results discussed above and presented in Fig. 3, a number of conclusions can be drawn regarding the contribution of expert predictor variables to the estimation and forecasting model performance. The first stage of adding expert variables considers the addition of decomposition variables (see Table 2 for specifications), the impact of including these variables is reflected by the performance difference between model configurations **1** and **2**, i.e. the models with subscript **1** and **2**. The performance improvement in terms of the MAE and RMSE is notable but gradual for the MLR and RF forecast models at both locations. For example, as a result of including the decomposition variables in MLR_2, the MAE reduces from 0.098 for MLR_1 to 0.093 kW/kWp at Bolzano. Such improvement is present to a lesser extent for the estimation models. In particular, the MLR estimation model applied to de Bilt does not profit from the added variables in configuration **2**. The contribution of the transposition variables, presented in Table 2, is captured in the performance difference between models **2** and **3**. The contribution of these variables to the estimation and forecasting models is significant for all models at both locations. For example, the MAE of the PV power output forecast model MLR

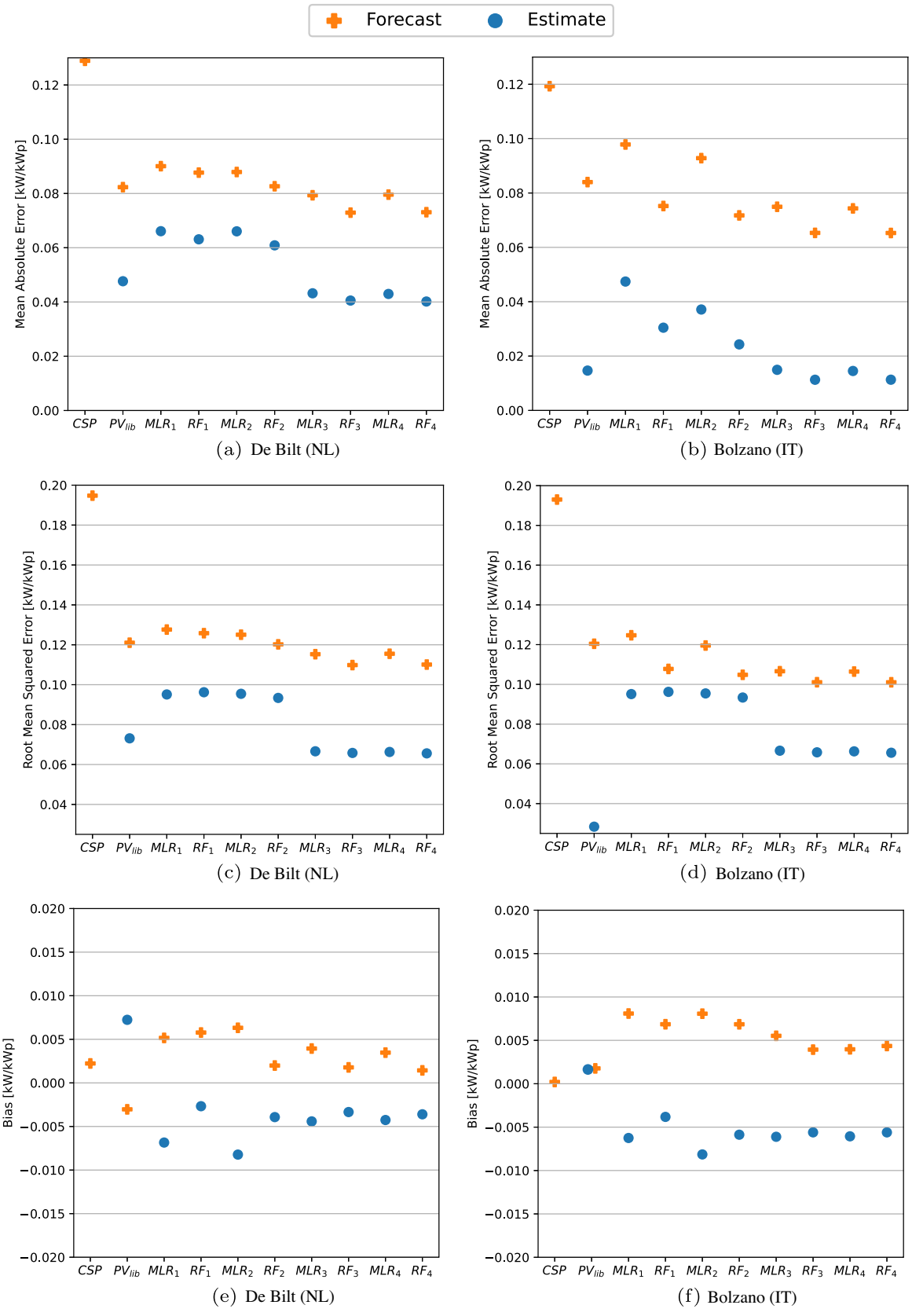


Fig. 3. Performance of the single-point estimation and forecasting models, expressed in the MAE (a,b), RMSE (c,d) and MBE (e,f) for de Bilt (a,c,e) and Bolzano (b,d,f). Different model configurations for incorporating expert variables are considered (see Table 2).

applied at Bolzano reduces from 0.093 for MLR_2 to 0.075 kW/kWp for MLR_3 (see Fig. 3(b)). Lastly, the inclusion of PV model variables in MLR_4 does not significantly improve the estimation or forecasting performance at both locations. For instance, an improvement of 0.001 in MAE was found for the MLR forecast model at Bolzano. Overall, the transposition variables have the most significant (positive) effect to the model performances.

Nonlinear and machine learning models such as RF used for PV power estimation and forecasting are in many studies found to significantly outperform MLR models (Hong et al., 2020; AlSkaif et al., 2020; Visser et al., 2022b). Similar results are found in this study, where model configurations RF_1 to RF_4 outperform their MLR counterparts in terms of the MAE and RMSE. Nevertheless, the results presented in Fig. 3 show that the integration of the decomposition and transposition model variables is relatively more valuable than using a machine learning model. This is substantiated by the performance results obtained for the MLR_3 estimation and forecasting models compared to RF_1 at both locations.

4.1.2. Probabilistic models

The performance of the probabilistic estimations and forecasting models is summarized in Fig. 4. Since we consider a prediction interval of 99% we expect a similar PICP score for each model. This is true for all models except the $CSPE$ forecast model at both locations and the QR_1 and QR_2 forecast models at Bolzano. The results show that most PICP scores range between 0.98 and 0.99. Interestingly, the PICP of the QRF_3 and QRF_4 forecast models are higher than the QRF_3 and QRF_4 estimation models. However, this is explained by a higher prediction interval, i.e. a higher PINAW value (see Figs. 4(c) and 4(d)). The PINAW scores range between 0.39 and 0.55 kW/kWp for the forecast models. The only exception is the $CSPE$ model, where a relative low PINAW score is found of 0.28 kW/kWp. The overall performance of the probabilistic models is best captured by the CRPS. The CRPS of the forecast models range between 0.04 and 0.08 kW/kWp. For de Bilt, all forecast models are found to outperform the benchmark model $CSPE$ substantially. At Bolzano the QR_1 performs similar to the $CSPE$ model. The CRPS of the estimation models is significantly better and varies between 0.008 and 0.0045 kW/kWp.

In general, the CRPS values of the probabilistic models follow the same trend as the MAE scores obtained for the single-point models. Remember, the CRPS reduces to the MAE for single-point values. The introduction of the variables produced by the decomposition model improves the performance of all models at both locations. In contrast to single-point models, this also covers the estimation models' performance. The value of the decomposition model is demonstrated by the performance differences between model configurations **1** and **2**. For example, at Bolzano the CRPS of the QRF model improves from 0.076 for QR_1 to 0.067 kW/kWp for QR_2. The value of including the transposition model is reflected in the performance differences of models **2** and **3**. Overall, the contribution of the transposition variables is larger compared to the decomposition variables. For instance, the QR model improves by 0.015 kW/kWp. Similar to the single-point models, the contribution of the PV model is limited. The CRPS of the QR forecast model at Bolzano decreases by less than 0.001 kW/kWp as a result of including the PV model variables.

Similar to the single-point model performance evaluation, it can be observed from Fig. 4(e) and 4(f) that the machine learning models are superior to regression ones. This is evident by comparing the estimation and forecasting results for QRF_1 to QRF_4 to their QR counterparts. Similarly, the inclusion of the decomposition and transposition variables in the QR model is more valuable than implementing the QRF model. This observation is identical to the results for the single-point models except that the QRF_1 forecast model at Bolzano is now found to outperform the QR_3 model in terms of the CRPS.

4.2. Contribution of expert variables

4.2.1. Correlations

A simple explanation for the improved performance as a result of including the expert variables created by the decomposition, transposition and PV models concerns the correlation between the constructed variables and the PV power output. Accordingly, the three main variables that are typically of interest to PV power estimation and forecasting applications because of their high correlation to the PV power output are presented in Fig. 5. The sub-figures present scatter plots of the predictor variables G_{GHI}, $G_{A.GI}$ and P_{PV} and the PV power output. In addition, the Pearson correlation coefficients are included on the top left corner of the figures. From Fig. 5 it is evident that the correlation with the registered PV power output, and therewith the linear relationship between the predictor and target variable, is significantly higher for $G_{A.GI}$ (e.g 0.95 for estimates at de Bilt) than G_{GHI} (0.90 for estimates at de Bilt). This holds for estimates and forecasts at both locations, explaining the increased performance of the model configuration **3** compared to configurations **1** and **2**. The correlation between the PV power output and P_{PV} is only slightly higher. This also explains the limited or absence of the model performance improvement as observed in Fig. 3 and 4 for model configuration **3** to **4**.

Another clear observation from Fig. 5 is the across the board increased correlation for estimates at Bolzano compared to de Bilt. This difference is due to the quality of G_{GHI} recordings at both locations. G_{GHI} measurements at Bolzano feature a higher temporal resolution and were in contrast to de Bilt recorded on the same site as where the PV system is located. The difference in correlation between forecasts at both locations is limited but is likely the result of more accurate weather forecasts at Bolzano. This difference can also be observed in Fig. 3 and 4.

4.2.2. Model dynamics

A better understanding of the improved model performance for the model configurations **3** and **4** can be observed from plotting the autocorrelation function (ACF) of the residuals. Since the ACF of residuals gives insights into the white noise properties of the residual, it can reveal model deficiencies and help identify missing elements in the model (Hyndman and Athanasopoulos, 2018; Martin et al., 2017) that might be overcome by introducing a single additional variable (Bacher and Madsen, 2011). Since the ACF relies on a complete time series, nighttime values were filled with zeros for the purpose of this evaluation. The ACF of the residuals for the single-point forecast models at de Bilt are presented in Fig. 6. The sub-figures present the lag (in hours) correlation of the residuals for all forecast horizons. Since the forecast time horizon varies between 12 and 36 hours-ahead, information captured in the lags up to 12 h is unavailable and availability for lags between 12 and 36 h is limited (plotted in gray). The ACF plots show a high lag dependency around lag values of 24, 48 and 76 h, for the MLR_1 and RF_1 models. This characterizes the typical diurnal pattern that defines the PV power output. The introduction of the decomposition variables improves the dynamics of the models, but fails to reach the assumption of white noise properties as the sub-figures still show a high lag correlation. The introduction of the transposition variables resolves the high lag correlation, i.e. the models now do not reject the assumption of white noise. This indicates that the forecast models MLR_3 and RF_3 are capable of describing the time-dependent dynamics of the power output of a PV system. The introduction of the PV model variables show little to no further improvement. In conclusion, this means that model configurations **3** and **4** are capable of describing the seasonality. Note, however, although these models now adequately describe the dynamics in the PV power output, the performance of these models may still be further improved. Similar results were found for the other single-point estimation and forecasting models tested in this study.

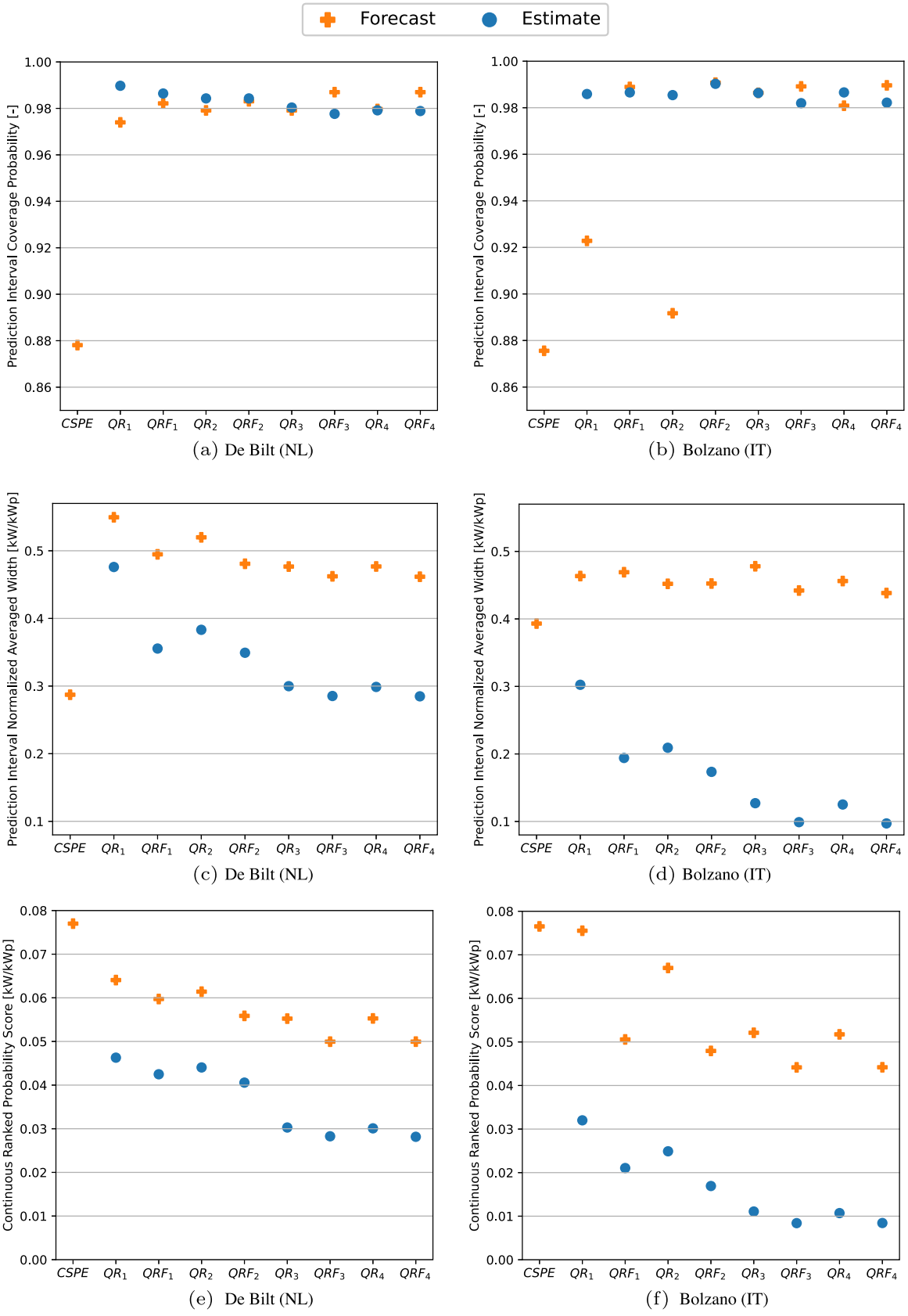
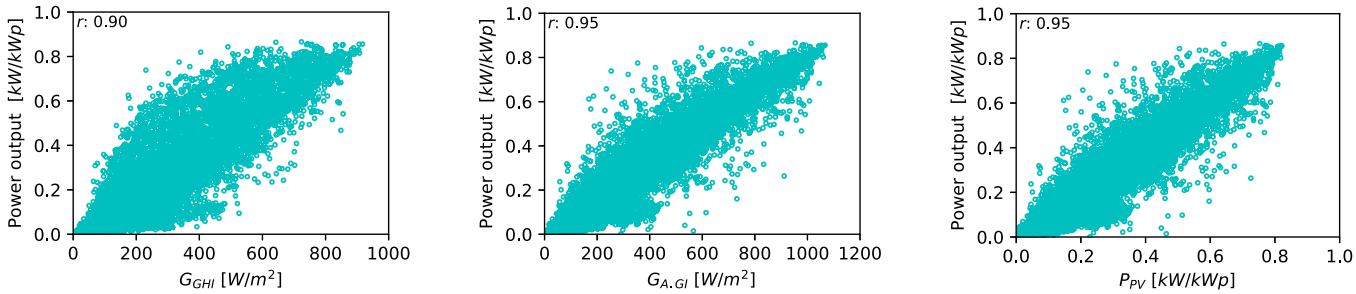
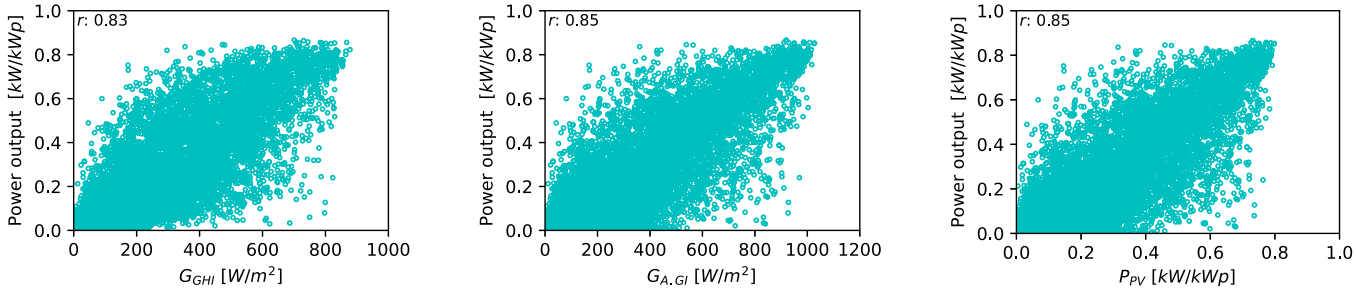


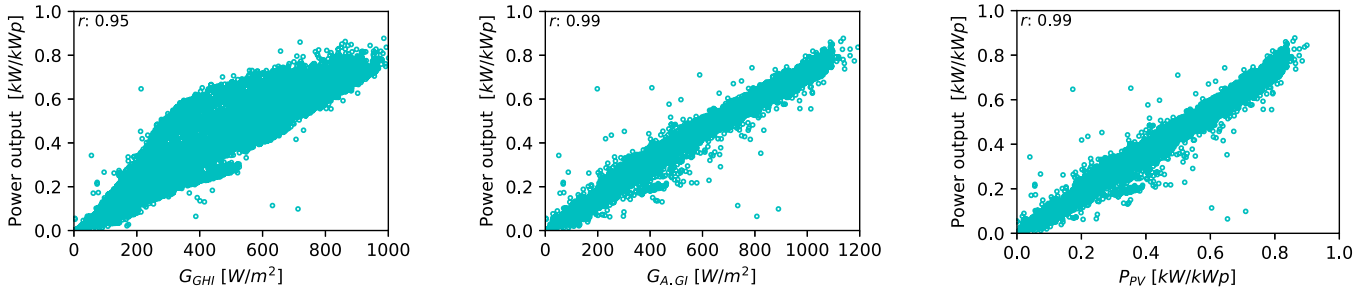
Fig. 4. Performance of the probabilistic PV power output estimation and forecasting models, expressed in the PICP (a,b), PINAW (c,d) and CRPS (e,f) for de Bilt (a,c,e) and Bolzano (b,d,f). Different model configurations for incorporating expert variables are considered (see Table 2).



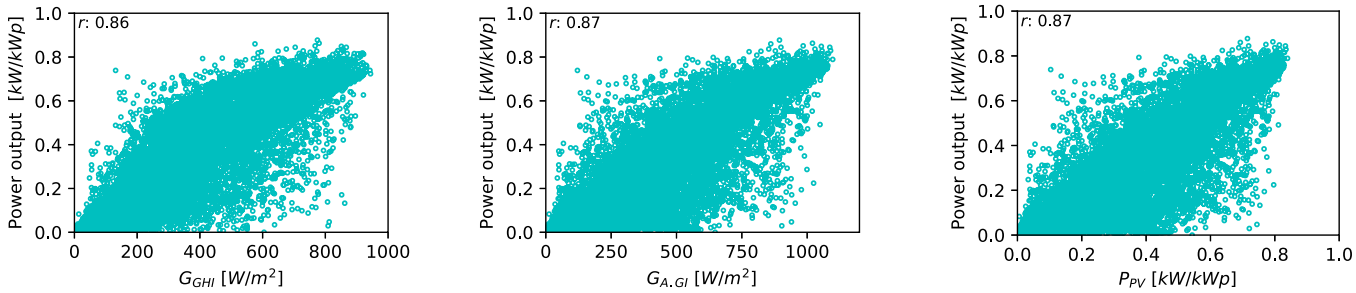
(a) Estimates, de Bilt (NL)



(b) Forecasts, de Bilt (NL)



(c) Estimates, Bolzano (IT)



(d) Forecasts, Bolzano (IT)

Fig. 5. Scatter plots between the predictor and target variable for estimating (a,c) and forecasting (b,d) the PV power output at De Bilt, Netherlands (a,b) and Bolzano, Italy (c,d). The plots are complemented with the Pearson correlation values (r) between predictor and target variable on the top left corner.

4.3. Selection of predictor variables

4.3.1. Single-point models

This section evaluates the optimal selection of predictor variables per model and case study by means of forward feature selection. These results provide insights into the most valuable predictor variable(s) and the ultimate set of predictor variables per model. The results for the single-point models are presented in Fig. 7. The figure shows that an optimal model performance for estimation models is reached at 3–5 predictor variables, i.e. the accuracy does not improve further after adding another variable. The forecasting models require more, namely 7–12 predictor variables to reach the optimal model performance. This

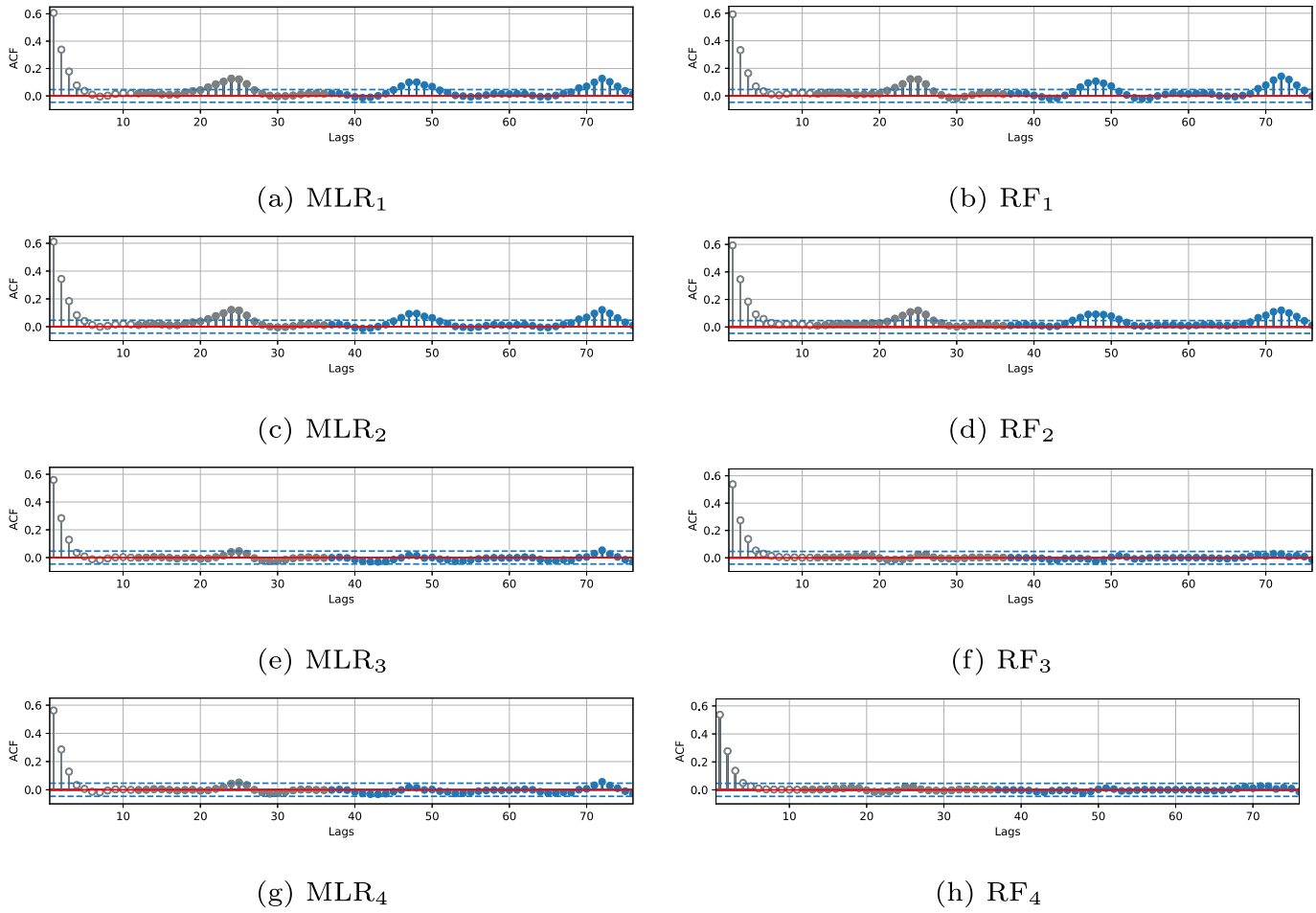


Fig. 6. Autocorrelation for single-point model forecasts at de Bilt, similar findings were obtained for the other single-point estimation and forecasting model applications.

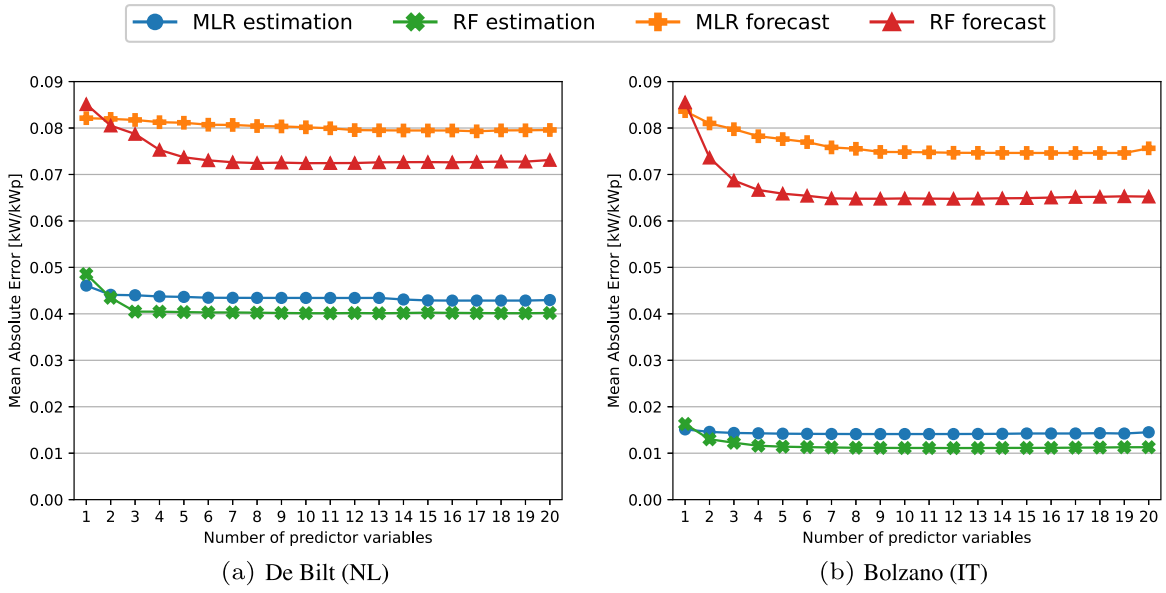


Fig. 7. Single-point estimation and forecasting model performance at de Bilt (a) and Bolzano (b) expressed as MAE to the number of variables selected according to the ranking presented in Tables 3 and B.1.

Table 4
The top five most important predictor variables for the probabilistic estimation and forecasting models at de Bilt and Bolzano. The most important variable is ranked No. 1. The values overlap with the order of included predictor variables as depicted on the x-axis in Fig. 8. A complete overview is given in Table B.2 in Appendix B.

No.	Estimation				Forecasting			
	De Bilt, NL		Bolzano, IT		De Bilt, NL		Bolzano, IT	
	QR	QRF	QR	QRF	QR	QRF	QR	QRF
#1	P_{PV}	P_{PV}	P_{PV}	P_{PV}	P_{PV}	P_{PV}	P_{PV}	P_{PV}
#2	G_{GHI}	G_{DNI}	$G_{A.BI}$	AM_A	$G_{A.BI}$	G_{DNI}	$G_{A.BI}$	AM_R
#3	$G_{A.BI}$	AM_A	G_{GHI}	T_D	G_{DNI}	AM_A	G_{DNI}	CC_T
#4	G_{DNI}	CC_T	P_A	G_{DNI}	G_{GHI}	CC_T	G_{GHI}	P_{MSL}
#5	WS_{v10}	WS_{u10}	G_{DHI}	CC_T	P_A	P_A	G_{DHI}	$G_{A.DI}$

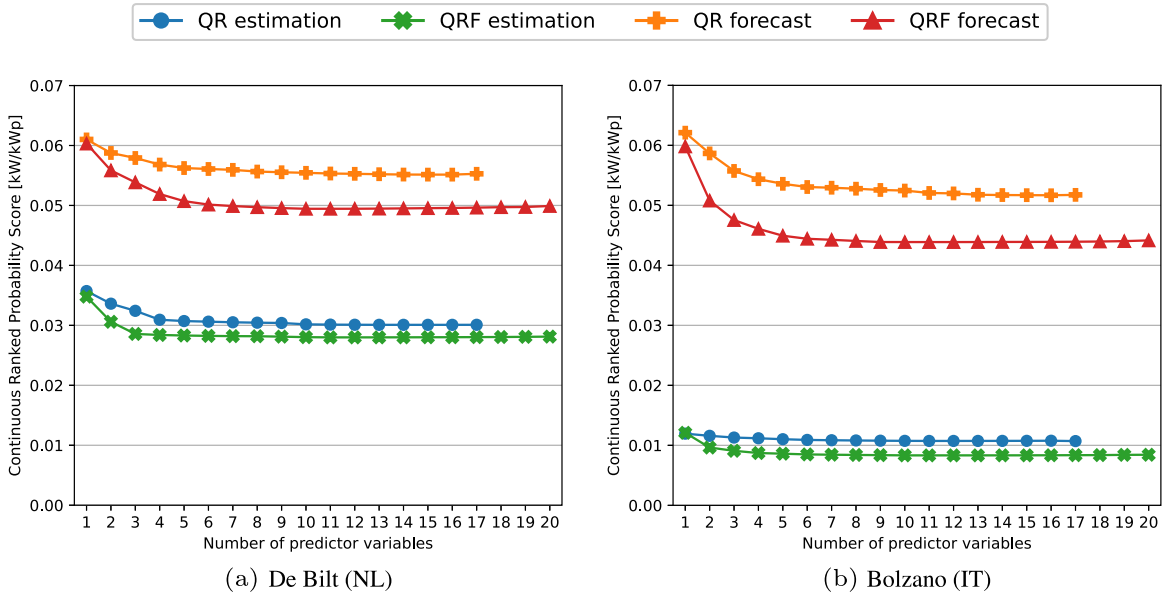


Fig. 8. Probabilistic estimation and forecasting model performance at de Bilt (a) and Bolzano (b) expressed as CRPS to the number of variables selected according to the ranking presented in Tables 4 and B.2. Note the limited number of predictor variables in case of the QR models, which is due to a multicollinearity issue (see Section 3.3).

difference is explained by the inherent uncertainty in the predictor variables used for forecasting, which translates into a larger number of predictor variables needed to reach optimal model performance.

The forward feature selection approach selects the single most important predictor variable, and then step by step adds the next most important predictor variable. The top five most important predictor variables for all models, and both case studies, is presented in Table 3 and a complete overview is given in Table B.1 in Appendix B. When studying these results, it is essential to consider that sometimes the differences between selecting two variables is insignificant, but due to high cross-correlation this could result in a predictor variable to be considered less valuable. An example is the variable G_{GHI} in the MLR forecasting model at de Bilt, which is ranked as the tenth variable. Nevertheless, its correlation with the fourth ranked variable $G_{A.GI}$ is extremely high, see Fig. 1. As a result, G_{GHI} would likely receive a higher ranking in case $G_{A.GI}$ was unavailable. P_{PV} is found to be the most important predictor variable in all models. This is in line with the results presented in Section 4.2.1, where the P_{PV} is shown to have the highest correlation with the target variable.

Another observation that stands out in Fig. 7 is the performance of the MLR models compared to their RF counterparts. The results once again show the superiority of the machine learning model over the MLR model, except for the case where only one predictor variable is considered. This holds for the single-point models at both locations. The superiority of the RF models is explained by their nonlinear nature.

4.3.2. Probabilistic models

The results for the probabilistic models correspond to the findings presented for the single-point models. The probabilistic estimation models require 3–5 predictor variables to achieve an optimal performance, see Fig. 8. In case of forecasting, a total of 6–11 predictor variables are required depending on the model and case study. Thus, the probabilistic forecasting models require a larger set of predictor variables, which is needed to describe the uncertainty in the forecasts. Similar to the single-point models, P_{PV} was found to be the most important predictor variable for all the estimation and forecasting models in both case studies. The top five most important predictor variables for all probabilistic models is given in Table 4, Table B.2 presents a complete overview of the ranking of all predictor variables. Finally, regarding the probabilistic models, the machine learning model (QRF) outperforms the linear regression model (QR) in estimation and forecasting for all sets of predictor variables at both case studies.

4.4. Sensitivity analysis

4.4.1. Single-point models

The results of the sensitivity analysis show how the tilt angle of the PV system affects the value of the expert variables. Fig. 9 presents the model performance in MAE as a function of the tilt angle, for the single-point models at both case studies. From Fig. 9 a couple of trends can

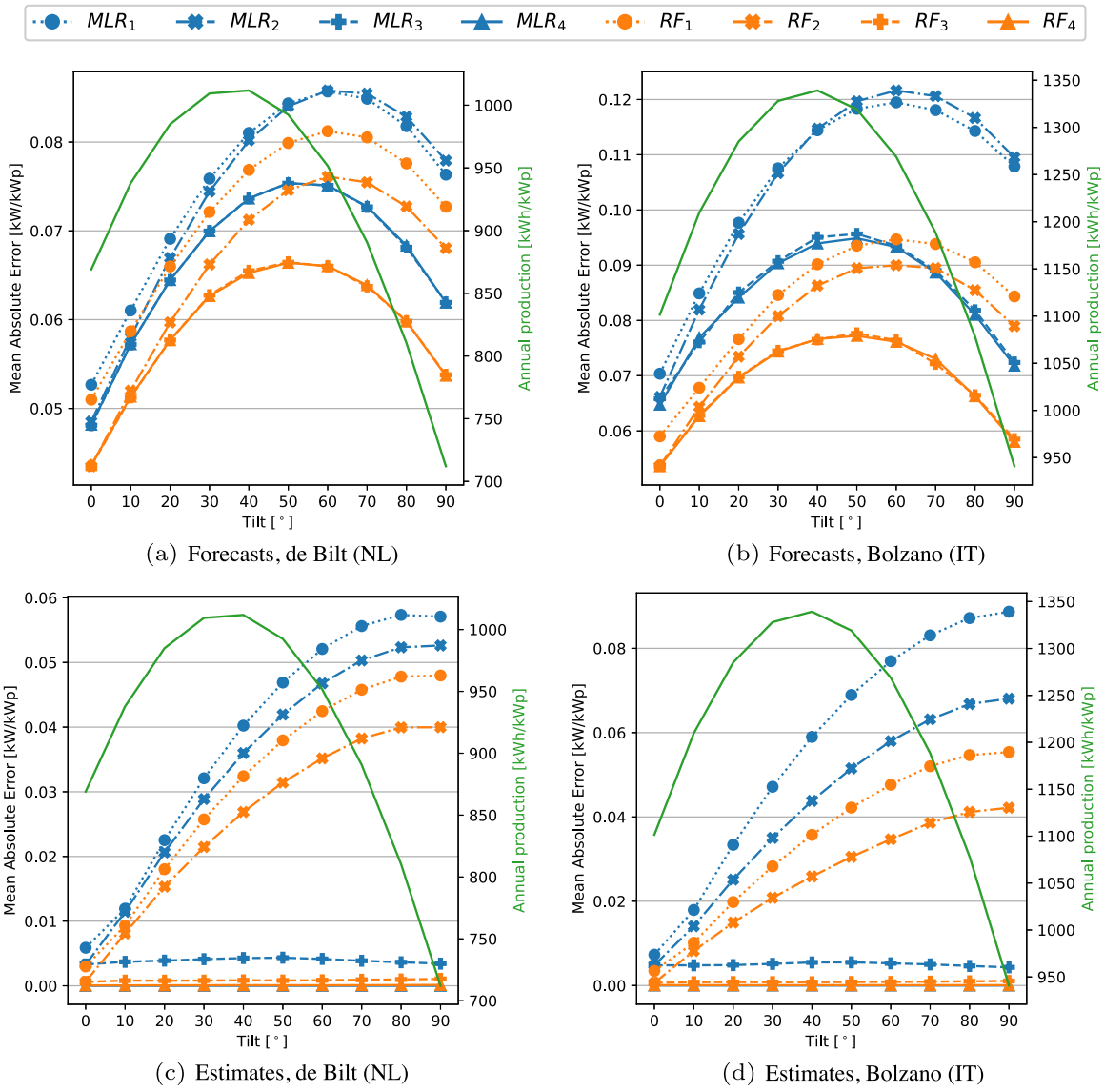


Fig. 9. Tilt dependency of the single-point estimation (c,d) and forecasting (a,b) models for de Bilt (a,c) and Bolzano (b,d). The performance is expressed in terms of the MAE. Note the different y-axis values.

be distinguished. As the results only differ in terms of the magnitude of the MAE per location, these trends can be generalized to both case studies. The sub-figures that present the results for the forecast models show a parabolic trend for all models. The performance of each forecast model deteriorates as the tilt angle increases up to $50-60°$, after which the performance improves. This general parabolic trend is observed for all forecast models, including configurations **3** and **4**. Since the expert variables are expected to correct for the influence of the tilt angle, the main driver of this parabolic trend is found in the specific yield of the PV system. Fig. 10 shows how the performance of the model differs when we correct the MAE for the specific yield, i.e. express the MAE in terms of kW/MWh on an annual basis. These results show a positive linear relationship between the forecast model performance and the tilt angle, where a larger slope is observed for the forecast models **1** and **2**. For all forecast models, this indicates a growing model forecast uncertainty with an increasing tilt angle.

The difference between the MAE values of model configurations **1** and **2** in Fig. 9 marks the value of the decomposition variables, which is significant for most tilt angles. The value of including the transposition variables in the estimation models is characterized by comparing the models **2** and **3**. Its value is considerable for tilt angles that exceed 10° and increases sharply with an increasing tilt angle. By comparing the model configurations **3** and **4** we find the contribution of the PV model variables, which is significant for the MLR model and marginal for the RF model.

4.4.2. Probabilistic models

The results of the sensitivity analysis for the probabilistic estimation and forecasting models are summarized in Fig. C1 and C2, see Appendix C. By comparing these to Fig. 9 and 10 similar trends are found in terms of the tilt dependency of the single-point and probabilistic estimation and forecasting models. The main difference concerns the extent of the errors, where lower error values are observed for the probabilistic models. Since the difference between the single-point and probabilistic models is limited to the magnitude of the error metric, it

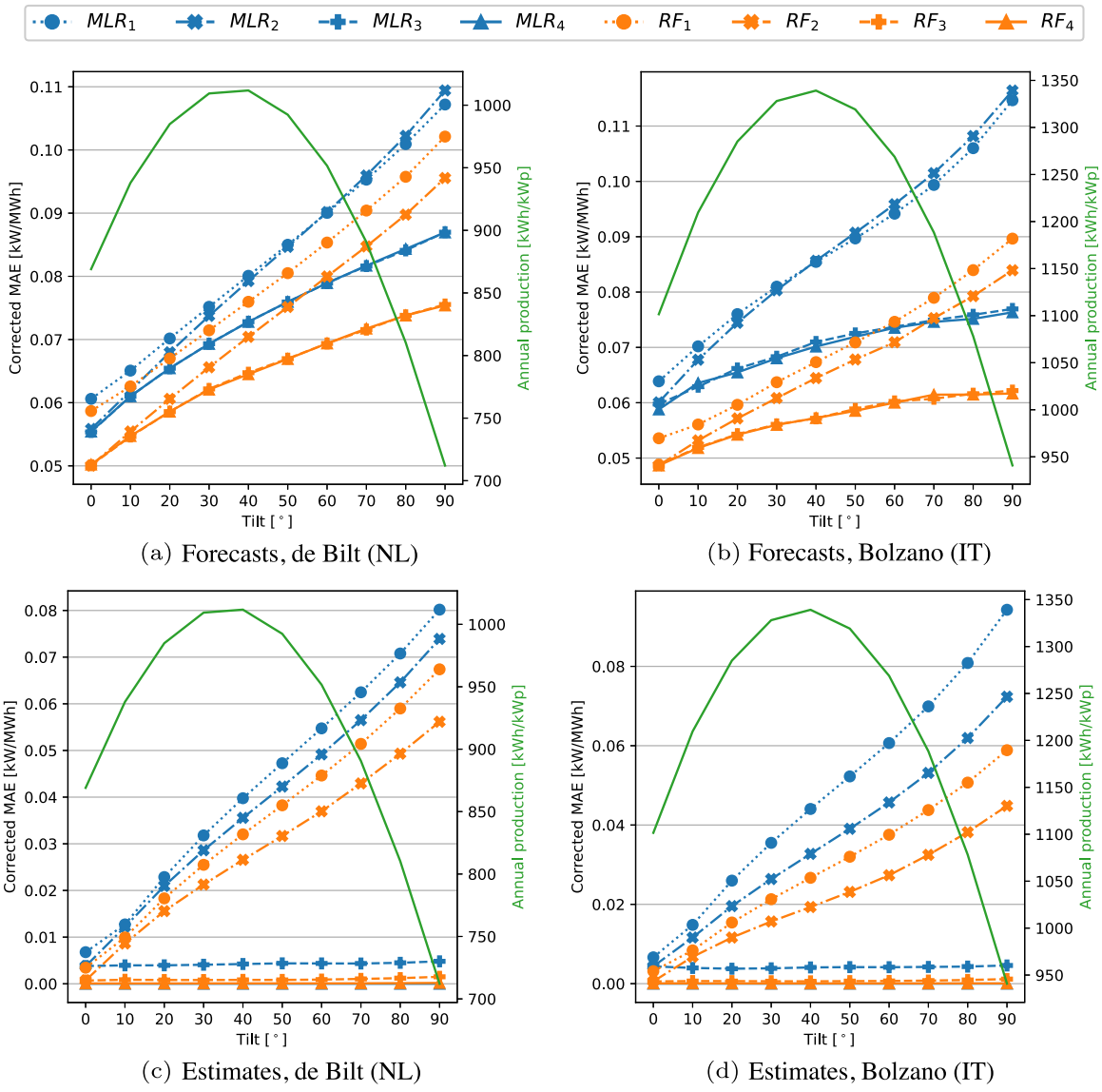


Fig. 10. Tilt dependency of the single-point estimation (c,d) and forecasting (a,b) models for de Bilt (a,c) and Bolzano (b,d). The performance is expressed in terms of the MAE corrected for the specific yield, i.e. kW/MWh on an annual basis. Note the different y-axis values.

can be concluded that the impact of the tilt angle on the model performance does not depend on the type of model of interest, i.e. single-point or probabilistic.

5. Conclusions

This study evaluates the value of predictor variables for the purpose of PV power estimation and forecasting. In particular, it quantifies the contribution of introducing expert variables in regression and machine learning models, which can easily be created using a decomposition, transposition and physical PV model in separate pre-processing steps. The results show the performance improvement of introducing each of these steps to single-point and probabilistic regression and machine learning models that estimate and forecast the PV power output for two different case studies. The performance of the estimation and forecasting models improve significantly after introducing the expert variables generated by the decomposition model. The variables generated by the transposition model are even more valuable. A closer look into the residuals reveal that the introduction of the variables outputted by the transposition model in particular improves the ability of the models to describe the diurnal dynamics a PV system is exposed to. In addition, per model and case study, the study identifies the optimal set of predictor variables. The forecasting models are found to require a larger set of predictor variables compared to the estimation models in order to describe the uncertainty. A sensitivity analysis shows the dependence of the contribution of expert variables for different tilt angles, whose value is significant for all tilt angles in case of the decomposition model and for tilt angles larger than 0° for the transposition model. Lastly, the results in this study demonstrate the superiority of the RF and QRF models compared to the MLR and QR models. Yet, the results also indicate that the inclusion of expert variables can be more valuable compared to using a more advanced model.

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.

Data availability statement

Weather forecasts and measurements are online available at https://www.ecmwf.int/en/forecasts/datasets/archive-datasets(ECMWF, 2020) and https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=overview (Muñoz Sabater, 2019). G_{GHI} measurements for de Bilt can be retrieved from https://projects.knmi.nl/klimatologie/uurgegevens/selectie.cgi (KNMI, 2020). PV power measurements for de Bilt can be found at https://doi.org/10.5281/zenodo.6906504 (Visser et al., 2022c). G_{GHI} and PV power measurements for Bolzano are confidential.

Acknowledgments

This work is part of the Energy Intranets (NEAT: ESI-BiDa 647.003.002) project, which is funded by the Dutch Research Council NWO in the framework of the Energy Systems Integration & Big Data programme.

Appendix A. Statistical properties

Similar to Figs. 1, A1, A3 and A2 hold the cross-correlation values of the predictor variables included in the measurements matrices for de Bilt and Bolzano, and the forecasting matrix for Bolzano. The depicted cross-correlation values are computed using the Pearson correlation coefficient between each pair of predictor variables. Although some differences are present, the cross-correlation values are very similar for both locations. The cross-correlation values found for the measurements matrices per location are almost identical to the forecast matrices.

The statistical properties of all matrices are summarized in Tables A.1 and A.2. Amongst others, these explain the higher annual yield of a PV system in Bolzano compared to de Bilt.

By comparing the results for the single-point forecast model configurations **1** and **2** as presented in Fig. 9 and 10, we can identify the tilt dependency of the value of including the decomposition variables. For the MLR model, this value is most significant for small tilt angles, i.e. $0-40°$. Alternatively, the performance improvement for the RF model is large for all tilt angles. The contribution of the transposition variables is presented by the performance difference between model configurations **2** and **3**. The value of the transposition variables is significant for both models at a tilt angle of at least $10°$, where a larger tilt angle increases the value of including these transposition variables. Since hardly any difference is found between the MAE values of the model configurations **3** and **4**, we can conclude that the PV model variables do not improve the performance of the single-point forecast models, regardless of the tilt angle of interest.

The results of the estimation models deviate slightly from the forecasting models. Fig. 9 shows that the performance of model configurations **1** and **2** deteriorate rapidly with an increasing tilt angle. Instead of a parabolic relation as was found for the single-point forecast models, with an increasing tilt the MAE increases linearly at first, and then flattens. Similar to the single-point forecast models, for estimation model configurations **1** and **2** we find a positive linear relation between the MAE and the tilt angle when the MAE is corrected for the specific yield. The MAE values of the model configurations **3** and **4** remain constant, and thus their performance is indifferent to the tilt angle (see Fig. 9 and 10). Note that model configuration **4** is fed with the true value, i.e. y is equal to P_{PV}. This also explains the good performance of model configuration **3**.

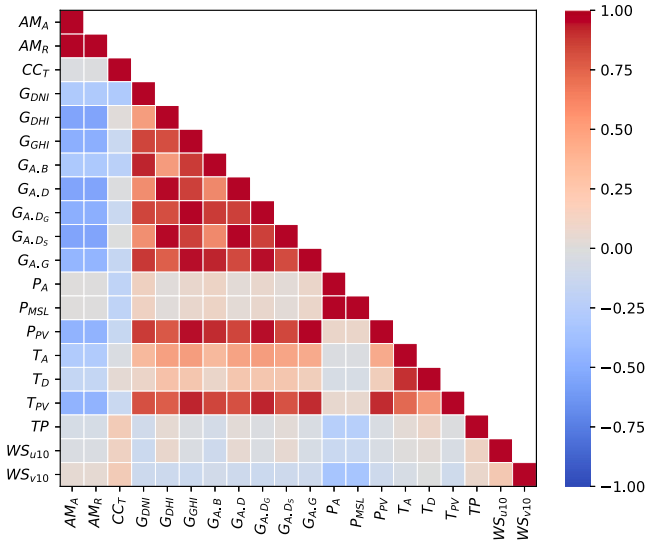


Fig. A1. Cross-correlation for the estimated values of the weather variables at de Bilt.

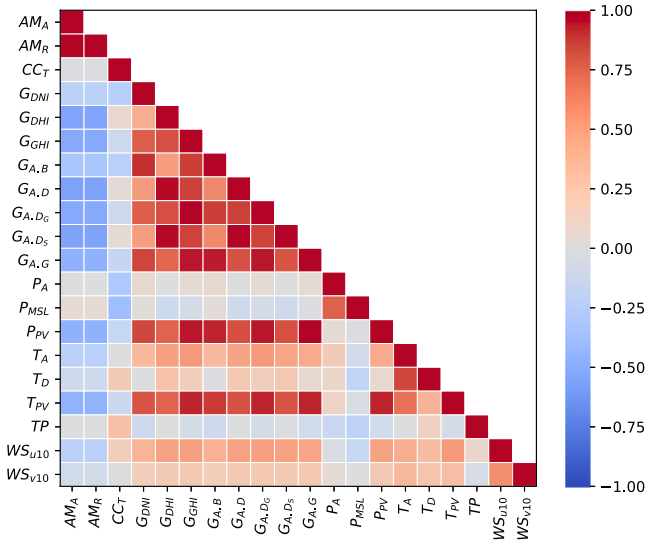


Fig. A2. Cross-correlation for the forecasted values of the weather variables at Bolzano.

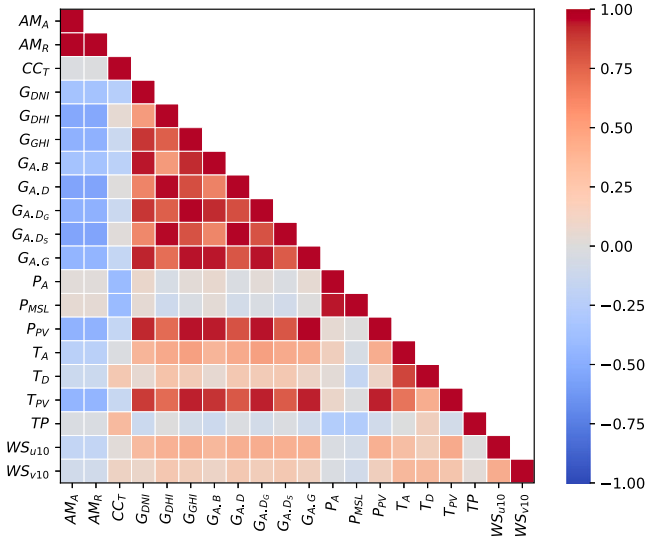


Fig. A3. Cross-correlation for the estimated values of the weather variables at Bolzano.

Table A.1
Statistical properties of the employed predictor variables, Measurements (M) and Forecasts (F) of the weather variables, for de Bilt.

	Mean		Median		Standard dev.		Minimum		Maximum		Unit
	M	F	M	F	M	F	M	F	M	F	
AM_A	0.05	0.05	0.05	0.05	0.04	0.04	0.01	0.01	0.38	0.38	[–]
AM_R	4.7	4.7	4.7	4.7	4.0	4.0	1.1	1.1	37.9	37.9	[–]
CC_T	0.67	0.65	0.82	0.82	0.35	0.37	0.0	0.0	1.0	1.0	[–]
G_{GHI}	123	125	5.1	4.9	190	192	0.0	0.0	887	878	W/m^2
G_{DNI}	112	117	0.00	0.00	221	227	0.00	0.00	1320	1255	W/m^2
G_{DHI}	70	69	5.1	4.9	99	97	0.00	0.00	395	394	W/m^2
$G_{A.GI}$	145	148	0.41	0.40	233	239	0.00	0.00	1045	1034	W/m^2
$G_{A.BI}$	70	74	0.00	0.00	153	160	0.00	0.00	855	842	W/m^2
$G_{A.DI}$	74	75	0.40	0.41	105	106	0.00	0.00	406	408	W/m^2
$G_{A.DI_g}$	2.4	2.4	0.10	0.09	3.6	3.7	0.00	0.00	17	17	W/m^2
$G_{A.DI_s}$	72	72	0.09	0.09	103	101	0.00	0.00	397	395	W/m^2
P_A	1014	1015	1015	1015	10	10	971	972	1044	1044	mbar
P_{MSL}	1015	1015	1016	1016	10	10	972	972	1045	1045	mbar
P_{PV}	287	293	0.00	0.00	461	470	0.00	0.00	1929	1913	W
T_A	11	11	11	11	6.2	6.2	−6.4	−7.0	33	33	°C
T_D	7.6	7.3	7.5	7.3	5.3	5.4	−8.3	−11	22	23	°C
T_{PV}	18	18	14	13	16	16	−6.4	−7.0	85	85	°C
TP	0.10	0.09	0.00	0.00	0.31	0.34	0.00	0.00	11	11	mm
WS_{u10}	0.94	1.1	1.0	1.2	2.8	3.0	−8.5	−9.8	12	13	m/s
WS_{v10}	0.91	1.0	1.1	1.2	2.7	2.9	−7.8	−8.9	11	12	m/s

Table A.2
Statistical properties of the employed predictor variables, Measurements (M) and Forecasts (F) of the weather variables, for Bolzano.

	Mean		Median		Standard dev.		Minimum		Maximum		Unit
	M	F	M	F	M	F	M	F	M	F	
AM_A	0.04	0.04	0.04	0.04	0.04	0.04	0.01	0.01	0.3	0.3	[–]
AM_R	4.3	4.3	4.35	4.35	4.02	4.02	1.09	1.09	38	38	[–]
CC_T	0.5	0.6	0.6	0.65	0.39	0.37	0.00	0.00	1.0	1.0	[–]
G_{DNI}	173	160	0.00	0.00	292	279	0.00	0.00	1482	1611	W/m^2
G_{DHI}	72	76	6.1	6.76	99	103	0.00	0.00	414	1632	W/m^2
G_{GHI}	159	155	6.1	6.79	230	223	0.00	0.00	944	1632	W/m^2
$G_{A.GI}$	198	190	0.4	0.43	288	278	0.00	0.00	1098	1128	W/m^2
$G_{A.BI}$	118	108	0.00	0.00	212	200	0.00	0.00	910	936	W/m^2
$G_{A.DI}$	80	82	0.4	0.43	107	111	0.00	0.00	421	421	W/m^2
$G_{A.DI_s}$	78	80	0.2	0.11	105	108	0.00	0.00	414	414	W/m^2
$G_{A.DI_g}$	1.9	1.9	0.07	0.08	2.78	2.69	0.00	0.00	11.38	20	W/m^2
P_A	909	906	908	906	9.2	7	867	867	938	924	mbar
P_{MSL}	1017	1017	1017	1017	7.5	7	976	978	1041	1040	mbar
P_{PV}	681	659	0.00	0.00	988	958	0	0	3526	3597	W
T_A	10	10	10	9.4	7.6	8	−10	−17	33	32	°C
T_D	3.1	3.2	3.6	3.7	7.4	7	−22	−23	20	19	°C
T_{PV}	21	21	14	14	21	20	−9.4	−17	89	88	°C
TP	0.1	0.1	0.00	0.00	0.4	0.4	0.00	0.00	14	6	mm
WS_{u10}	0.06	−0.08	0.00	−0.1	0.6	0.3	−2.3	−1.2	2.8	1.8	m/s
WS_{v10}	−0.12	−0.06	−0.1	−0.2	0.7	0.7	−4.6	−4.0	3.0	3.0	m/s

Appendix B. Ranking of predictor variables

Table B.1 and B.2 present the most important predictor variables for all single-point and probabilistic estimation and forecasting models at de Bilt and Bolzano.

Appendix C. Sensitivity

Fig. C1 and C2 present the probabilistic model performance for varying tilt angles, which relate to the results discussed in Section 4.4.2. The results are similar to the single-point models, as discussed in Section 4.4.1.

Table B.1
Overview of the ranking of the most important predictor variables for the single-point estimation and forecasting models at de Bilt and Bolzano. The most important variable is ranked No. 1. The values overlap with the order of included predictor variables as depicted on the x-axis in Fig. 7.

No.	Estimation				Forecasting			
	De Bilt, NL		Bolzano, IT		De Bilt, NL		Bolzano, IT	
	MLR	RF	MLR	RF	MLR	RF	MLR	RF
#1	P_{PV}	P_{PV}	P_{PV}	P_{PV}	P_{PV}	P_{PV}	P_{PV}	P_{PV}
#2	G_{GHI}	$G_{A.DI_G}$	G_{GHI}	G_{GHI}	WS_{u10}	$G_{A.DI_G}$	G_{GHI}	AM_R
#3	G_{DNI}	G_{DNI}	$G_{A.GI}$	G_{DNI}	G_{DNI}	G_{DNI}	G_{DNI}	CC_T
#4	$G_{A.GI}$	P_{MSL}	T_D	T_D	$G_{A.GI}$	CC_T	CC_T	P_{MSL}
#5	WS_{v10}	AM_A	G_{DHI}	CC_T	AM_A	TP	AM_A	$G_{A.DI_S}$
#6	WS_{u10}	WS_{u10}	AM_R	WS_{v10}	CC_T	P_A	P_{MSL}	WS_{u10}
#7	P_{MSL}	CC_T	WS_{v10}	AM_A	P_A	WS_{u10}	AM_R	TP
#8	P_A	$G_{A.DI}$	TP	TP	AM_R	$G_{A.BI}$	G_{DHI}	AM_A
#9	TP	P_A	CC_T	$G_{A.DI_S}$	P_{MSL}	$G_{A.GI}$	$G_{A.DI}$	G_{GHI}
#10	AM_A	WS_{v10}	AM_A	$G_{A.BI}$	G_{GHI}	T_{PV}	WS_{v10}	$G_{A.DI_G}$
#11	AM_R	$G_{A.DI_S}$	$G_{A.DI_G}$	AM_R	G_{DHI}	G_{GHI}	T_{PV}	$G_{A.DI}$
#12	$G_{A.DI_G}$	G_{DHI}	P_A	$G_{A.DI}$	$G_{A.DI_S}$	P_{MSL}	T_D	G_{DHI}
#13	T_{PV}	G_{GHI}	P_{MSL}	$G_{A.DI_G}$	WS_{v10}	T_A	$G_{A.GI}$	T_A
#14	T_D	AM_R	WS_{u10}	G_{DHI}	TP	T_D	T_A	T_{PV}
#15	T_A	T_{PV}	$G_{A.DI}$	WS_{u10}	$G_{A.DI}$	$G_{A.DI_S}$	WS_{u10}	$G_{A.GI}$
#16	G_{DHI}	T_D	$G_{A.DI_S}$	$Ga_{A.GI}$	$G_{A.DI_G}$	G_{DHI}	TP	T_D
#17	$G_{A.BI}$	T_A	$G_{A.BI}$	T_{PV}	$G_{A.BI}$	AM_R	$G_{A.DI_S}$	P_A
#18	$G_{A.DI}$	TP	T_A	T_A	T_{PV}	AM_A	$G_{A.BI}$	G_{DNI}
#19	$G_{A.DI_S}$	$G_{A.BI}$	T_{PV}	P_A	T_D	$G_{A.DI}$	$G_{A.DI_G}$	$G_{A.BI}$
#20	CC_T	$G_{A.GI}$	G_{DNI}	P_{MSL}	T_A	WS_{v10}	P_A	WS_{v10}

Table B.2
Overview of the ranking of the most important predictor variables for the probabilistic estimation and forecasting models at de Bilt and Bolzano. The most important variable is ranked No. 1. The values overlap with the order of included predictor variables as depicted on the x-axis in Fig. 8.

No.	Estimation				Forecasting			
	De Bilt, NL		Bolzano, IT		De Bilt, NL		Bolzano, IT	
	QR	QRF	QR	QRF	QR	QRF	QR	QRF
#1	P_{PV}	P_{PV}	P_{PV}	P_{PV}	P_{PV}	P_{PV}	P_{PV}	P_{PV}
#2	G_{GHI}	G_{DNI}	$G_{A.BI}$	AM_A	$G_{A.BI}$	G_{DNI}	$G_{A.BI}$	AM_R
#3	$G_{A.BI}$	AM_A	G_{GHI}	T_D	G_{DNI}	AM_A	G_{DNI}	CC_T
#4	G_{DNI}	CC_T	P_A	G_{DNI}	G_{GHI}	CC_T	G_{GHI}	P_{MSL}
#5	WS_{v10}	WS_{u10}	G_{DHI}	CC_T	P_A	P_A	G_{DHI}	$G_{A.DI}$
#6	G_{DHI}	P_A	T_D	G_{GHI}	AM_A	TP	$G_{A.GI}$	T_D
#7	$G_{A.GI}$	$G_{A.DI}$	G_{DNI}	WS_{v10}	G_{DHI}	WS_{u10}	CC_T	TP
#8	WS_{u10}	G_{GHI}	$G_{A.GI}$	TP	$G_{A.GI}$	$G_{A.BI}$	P_{MSL}	WS_{u10}
#9	T_D	T_D	TP	$G_{A.DI_S}$	P_{MSL}	T_{PV}	WS_{v10}	G_{GHI}
#10	T_A	T_A	AM_A	$G_{A.BI}$	CC_T	G_{GHI}	AM_A	AM_A
#11	AM_A	TP	CC_T	AM_R	AM_R	$G_{A.DI_G}$	AM_R	$G_{A.DI_G}$
#12	TP	$G_{A.DI_G}$	AM_R	$G_{A.DI_G}$	T_{PV}	P_{MSL}	T_{PV}	$G_{A.DI_S}$
#13	AM_R	$G_{A.DI_S}$	WS_{v10}	$G_{A.DI}$	T_D	T_D	T_A	G_{DHI}
#14	CC_T	AM_R	P_{MSL}	T_A	TP	$G_{A.GI}$	T_D	T_{PV}
#15	P_{MSL}	P_{MSL}	WS_{u10}	G_{DHI}	WS_{u10}	AM_R	TP	P_A
#16	T_{PV}	$G_{A.BI}$	T_{PV}	T_{PV}	WS_{v10}	T_A	WS_{u10}	G_{DNI}
#17	P_A	G_{DHI}	T_A	WS_{u10}	T_A	G_{DHI}	P_A	$G_{A.GI}$
#18		T_{PV}		$G_{A.GI}$		$G_{A.DI_S}$		T_A
#19		WS_{v10}		P_A		$G_{A.DI}$		$G_{A.BI}$
#20		$G_{A.GI}$		P_{MSL}		WS_{v10}		WS_{v10}

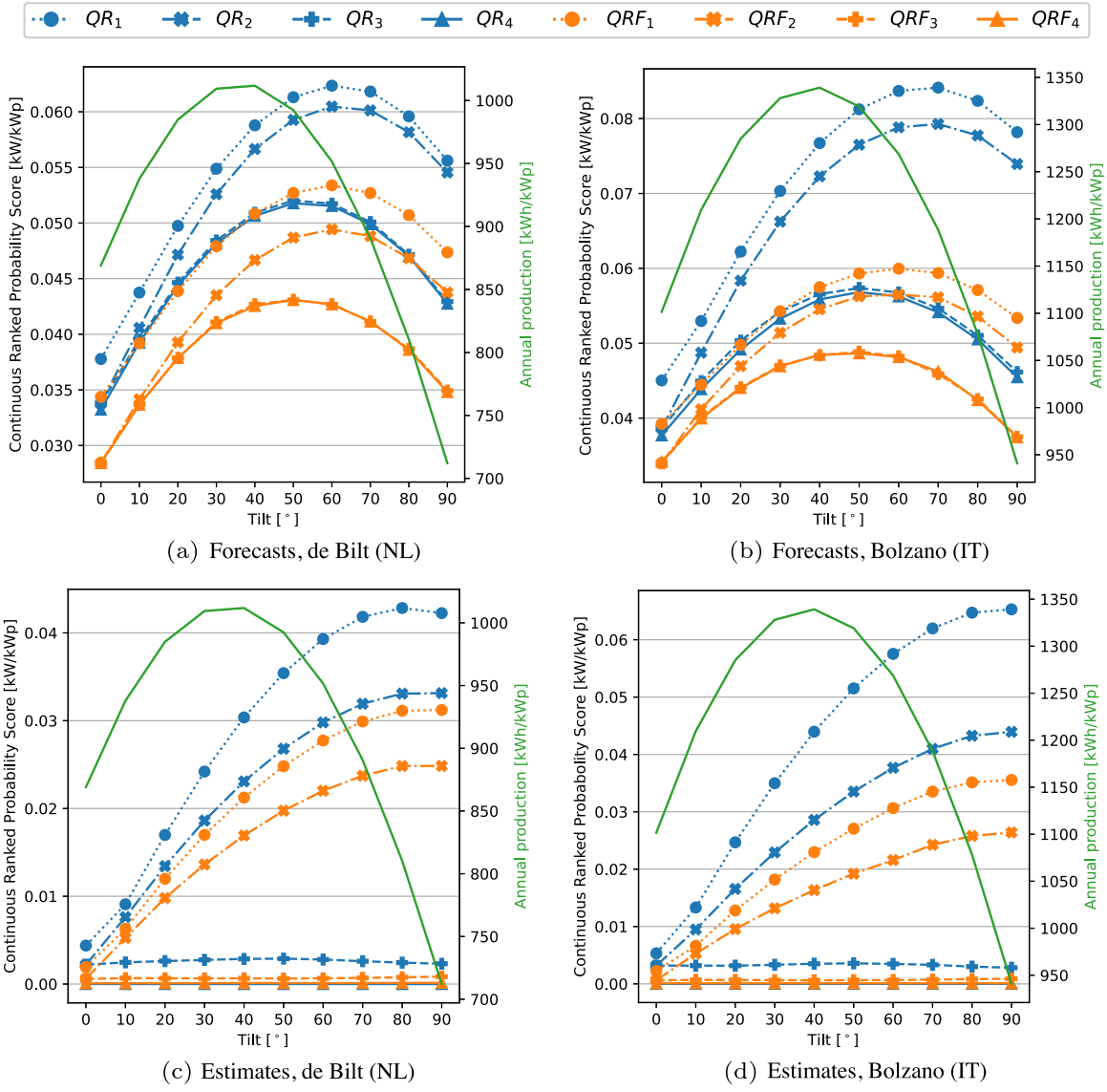


Fig. C1. Tilt dependency of the probabilistic estimation (c,d) and forecasting (a,b) models for de Bilt (a,c) and Bolzano (b,d). The performance is expressed in terms of the CRPS.

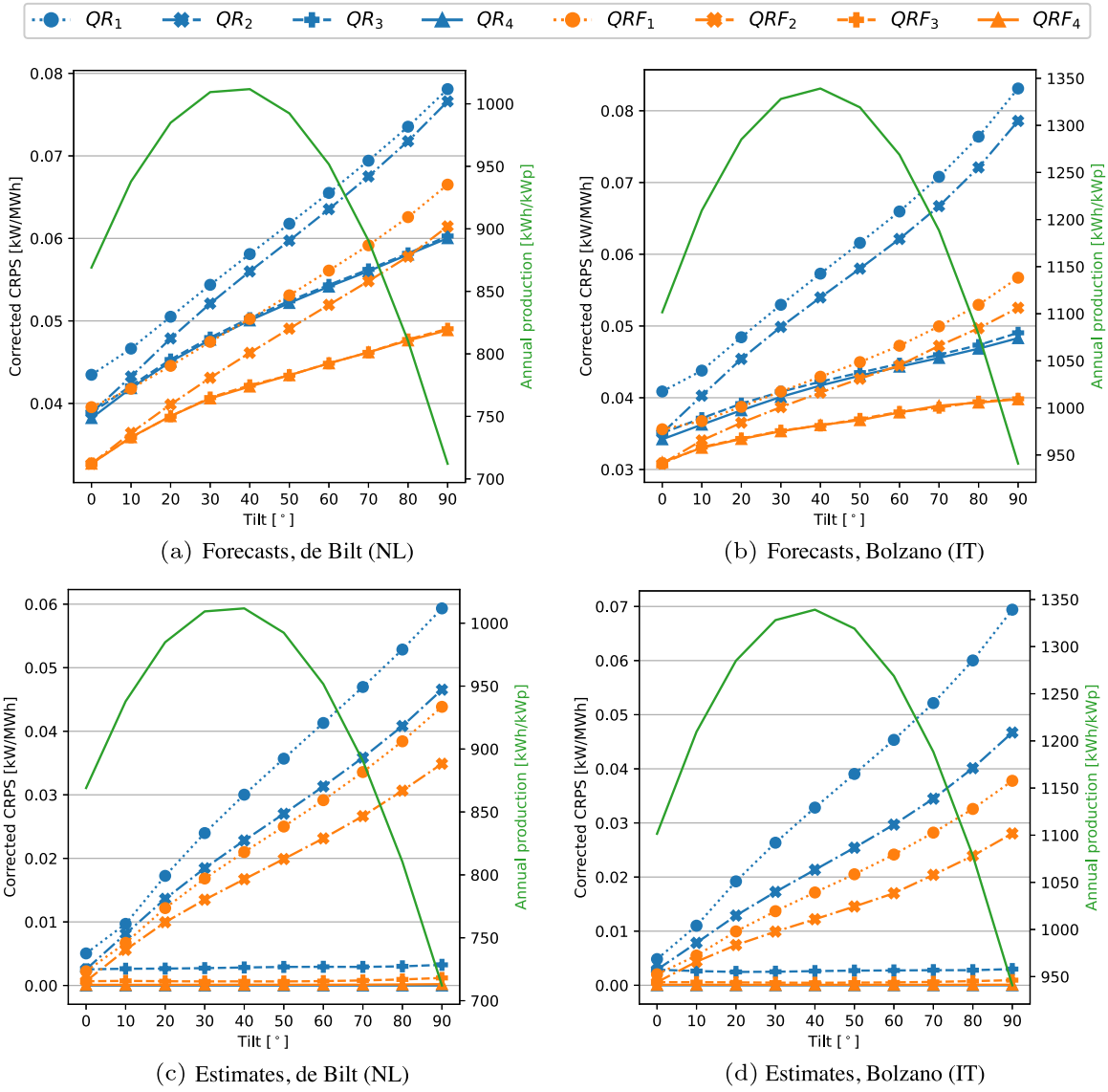


Fig. C2. Tilt dependency of the probabilistic estimation (c,d) and forecasting (a,b) models for de Bilt (a,c) and Bolzano (b,d). The performance is expressed in terms of the CRPS corrected for the specific yield, i.e. kW/MWh on an annual basis.

References

Ahmed, R., Sreeram, V., Mishra, Y., Arif, M., 2020. A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization. Renew. Sustain. Energy Rev. 124, 109792.

Akhter, M.N., Mekhilef, S., Mokhlis, H., Mohamed Shah, N., 2019. Review on forecasting of photovoltaic power generation based on machine learning and metaheuristic techniques. IET Renew. Power Gener. 13 (7), 1009–1023.

AlSkaif, T., Dev, S., Visser, L., Hossari, M., van Sark, W., 2020. A systematic analysis of meteorological variables for PV output power estimation. Renew. Energy 153, 12–22.

Altmann, A., Toloşi, L., Sander, O., Lengauer, T., 2010. Permutation importance: a corrected feature importance measure. Bioinformatics 26 (10), 1340–1347.

Bacher, P., Madsen, H., 2011. Identifying suitable models for the heat dynamics of buildings. Energy Build. 43 (7), 1511–1522.

Bemister-Buffington, J., Wolf, A.J., Raschka, S., Kuhn, L.A., 2020. Machine learning to identify flexibility signatures of class a GPCR inhibition. Biomolecules 10 (3), 454.

Boland, J., 2020. Characterising seasonality of solar radiation and solar farm output. Energies 13 (2), 471.

Boyson, W.E., Galbraith, G.M., King, D.L., Gonzalez, S., 2007. Performance model for grid-connected photovoltaic inverters. Technical Report, Sandia National Laboratories (SNL), Albuquerque, NM, and Livermore, CA.

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

Brinkel, N., Visser, L., AlSkaif, T., Van Sark, W., 2021. Avoiding low-voltage grid congestion using smart charging of electric vehicles based on day-ahead probabilistic photovoltaic forecasts. In: 2021 International Conference on Smart Energy Systems and Technologies. SEST, IEEE, pp. 1–6.

Díaz, G., Coto, J., Gómez-Aleixandre, J., 2019. Prediction and explanation of the formation of the Spanish day-ahead electricity price through machine learning regression. Appl. Energy 239, 610–625.

ECMWF, 2020. European centre for medium-range weather forecasts, ECMWF. URL https://www.ecmwf.int/en/forecasts/datasets/archive-datasets.

Erbs, D., Klein, S., Duffie, J., 1982. Estimation of the diffuse radiation fraction for hourly, daily and monthly-average global radiation. Sol. Energy 28 (4), 293–302.

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

Hong, T., Pinson, P., Wang, Y., Weron, R., Yang, D., Zareipour, H., 2020. Energy forecasting: A review and outlook. IEEE Open Access J. Power Energy 7, 376–388.

Hyndman, R.J., Athanasopoulos, G., 2018. Forecasting: Principles and Practice. OTexts.

IEA, 2020. World energy outlook 2020. Technical Report, IEA, Paris, p. 463, URL https://www.iea.org/reports/world-energy-outlook-2020.

Ineichen, P., Perez, R., 2002. A new airmass independent formulation for the Linke turbidity coefficient. Sol. Energy 73 (3), 151–157.

Jeon, J., Taylor, J.W., 2012. Using conditional kernel density estimation for wind power density forecasting. J. Amer. Statist. Assoc. 107 (497), 66–79.

Kasten, F., Young, A.T., 1989. Revised optical air mass tables and approximation formula. Appl. Opt. 28 (22), 4735–4738.

KNMI, 2020. Klimatologie: Uurgegevens van het weer in nederland. URL https://projects.knmi.nl/klimatologie/uurgegevens/selectie.cgi.

Koenker, R., Hallock, K.F., 2001. Quantile regression. J. Econ. Perspect. 15 (4), 143–156.

Kratochvil, J.A., Boyson, W.E., King, D.L., 2004. Photovoltaic array performance model. http://dx.doi.org/10.2172/919131, URL https://www.osti.gov/biblio/919131.

Lauret, P., David, M., Pedro, H.T., 2017. Probabilistic solar forecasting using quantile regression models. Energies 10 (10), 1591.

Lauret, P., David, M., Pinson, P., 2019. Verification of solar irradiance probabilistic forecasts. Sol. Energy 194, 254–271.

Lauret, P., Voyant, C., Soubdhan, T., David, M., Poggi, P., 2015. A benchmarking of machine learning techniques for solar radiation forecasting in an insular context. Sol. Energy 112, 446–457.

Louwen, A., Schropp, R.E., van Sark, W.G., Faaij, A.P., 2017. Geospatial analysis of the energy yield and environmental footprint of different photovoltaic module technologies. Sol. Energy 155, 1339–1353.

Markovics, D., Mayer, M.J., 2022. Comparison of machine learning methods for photovoltaic power forecasting based on numerical weather prediction. Renew. Sustain. Energy Rev. 161, 112364.

Martin, J., De Adana, D.D.R., Asuero, A.G., 2017. Fitting models to data: residual analysis, a primer. In: Uncertainty Quantification and Model Calibration. Vol. 133, Rijeka, HR: Intech.

Mayer, M.J., 2022. Benefits of physical and machine learning hybridization for photovoltaic power forecasting. Renew. Sustain. Energy Rev. 168, 112772.

Mayer, M.J., Gróf, G., 2021. Extensive comparison of physical models for photovoltaic power forecasting. Appl. Energy 283, 116239.

Meinshausen, N., Ridgeway, G., 2006. Quantile regression forests. J. Mach. Learn. Res. 7 (6).

Muñoz Sabater, J., 2019. ERA5-land hourly data from 1981 to present.

Nguyen, T.N., Müsgens, F., 2022. What drives the accuracy of PV output forecasts? Appl. Energy 323, 119603.

Pedro, H.T., Coimbra, C.F., David, M., Lauret, P., 2018. Assessment of machine learning techniques for deterministic and probabilistic intra-hour solar forecasts. Renew. Energy 123, 191–203.

Perez, R., Ineichen, P., Seals, R., Michalsky, J., Stewart, R., 1990. Modeling daylight availability and irradiance components from direct and global irradiance. Sol. Energy 44 (5), 271–289.

Perez, R., Perez, M., Schlemmer, J., Dise, J., Hoff, T.E., Swierc, A., Keelin, P., Pierro, M., Cornaro, C., 2020. From firm solar power forecasts to firm solar power generation an effective path to ultra-high renewable penetration a new york case study. Energies 13 (17), 4489.

Pombo, D.V., Bacher, P., Ziras, C., Bindner, H.W., Spataru, S.V., Sørensen, P.E., 2022. Benchmarking physics-informed machine learning-based short term pv-power forecasting tools. Energy Rep. 8, 6512–6520.

Rana, M., Rahman, A., 2020. Multiple steps ahead solar photovoltaic power forecasting based on univariate machine learning models and data re-sampling. Sustain. Energy Grids Netw. 21, 100286.

Raschka, S., Mirjalili, V., 2019. Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-Learn, and TensorFlow 2. Packt Publishing Ltd.

Reda, I., Andreas, A., 2004. Solar position algorithm for solar radiation applications. Sol. Energy 76 (5), 577–589.

Singla, P., Duhan, M., Saroha, S., 2021. A comprehensive review and analysis of solar forecasting techniques. Front. Energy 1–37.

Sobri, S., Koohi-Kamali, S., Rahim, N.A., 2018. Solar photovoltaic generation forecasting methods: A review. Energy Convers. Manage. 156, 459–497.

Spencer, J., 1982. A comparison of methods for estimating hourly diffuse solar radiation from global solar radiation. Sol. Energy 29 (1), 19–32.

Tripathy, D.S., Prusty, B.R., Jena, D., Sahu, M.K., 2020. Multi-time instant probabilistic PV generation forecasting using quantile regression forests. In: 2020 IEEE 9th Power India International Conference. PIICON, IEEE, pp. 1–6.

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

Visser, L., AlSkaif, T., van Sark, W., 2022a. An evaluation of predictor variables for photovoltaic power forecasting. In: International Conference on Intelligent Technologies and Applications. Springer, pp. 303–310.

Visser, L., AlSkaif, T., van Sark, W., 2022b. Operational day-ahead solar power forecasting for aggregated PV systems with a varying spatial distribution. Renew. Energy 183, 267–282.

Visser, L., AlSkaif, T., Van Sark, W., 2020. The importance of predictor variables and feature selection in day-ahead electricity price forecasting. In: 2020 International Conference on Smart Energy Systems and Technologies. SEST, IEEE, pp. 1–6.

Visser, L., Elsinga, B., AlSkaif, T., van Sark, W., 2022c. Open-source quality control routine and multi-year power generation data of 175 PV systems. http://dx.doi.org/10.5281/zenodo.6906504.

Visser, L.R., Elsinga, B., AlSkaif, T.A., Van Sark, W.G., 2022d. Open-source quality control routine and multi-year power generation data of 175 PV systems. J. Renew. Sustain. Energy 14 (4), 043501.

Visser, L.R., Lorenz, E., Heinemann, D., van Sark, W.G.J.H.M., 2022e. Solar power forecasts. In: Fthenakis, V., van Sark, W.G.J.H.M. (Eds.), Photovoltaic Technology, second ed. Elsevier, United Kingdom, pp. 213–233, Volume 1 in Comprehensive Renewable Energy, 2nd edition, T. Letcher (Ed.).

Xu, M., Pinson, P., Lu, Z., Qiao, Y., Min, Y., 2016. Adaptive robust polynomial regression for power curve modeling with application to wind power forecasting. Wind Energy 19 (12), 2321–2336.

Yang, D., van der Meer, D., 2021. Post-processing in solar forecasting: Ten overarching thinking tools. Renew. Sustain. Energy Rev. 140, 110735.

Yang, D., Wang, W., Gueymard, C.A., Hong, T., Kleissl, J., Huang, J., Perez, M.J., Perez, R., Bright, J.M., Xia, X., et al., 2022. A review of solar forecasting, its dependence on atmospheric sciences and implications for grid integration: Towards carbon neutrality. Renew. Sustain. Energy Rev. 161, 112348.

Young, P.C., Pedregal, D.J., Tych, W., 1999. Dynamic harmonic regression. J. Forecast. 18 (6), 369–394.