

# A systematic analysis of meteorological variables for PV output power estimation



Tarek AlSkaif<sup>a,\*,1</sup>, Soumyabrata Dev<sup>b,c,1</sup>, Lennard Visser<sup>a,1</sup>, Murhaf Hossari<sup>b</sup>, Wilfried van Sark<sup>a</sup>

<sup>a</sup> Copernicus Institute of Sustainable Development, Utrecht University, Utrecht, the Netherlands

<sup>b</sup> ADAPT SFI Research Centre, Dublin, Ireland

<sup>c</sup> School of Computer Science, University College Dublin, Ireland

## ARTICLE INFO

### Article history:

Received 13 August 2019

Received in revised form

17 December 2019

Accepted 31 January 2020

Available online 5 February 2020

### Keywords:

Photovoltaic

Solar power estimation

Meteorological variables

Machine learning

Regression methods

## ABSTRACT

While the large-scale deployment of photovoltaics (PV) for generating electricity plays an important role to mitigate global warming, the variability of PV output power poses challenges in grid management. Typically, the PV output power is dependent on various meteorological variables at the PV site. In this paper, we present a systematic approach to perform an analysis on different meteorological variables, namely temperature, dew point temperature, relative humidity, visibility, air pressure, wind speed, cloud cover, wind bearing and precipitation, and assess their impact on PV output power estimation. The study uses three years of input meteorological data and PV output power data from multiple prosumers in two case studies, one in the U.S. and one in the Netherlands. The analysis covers the correlation and interdependence among the meteorological variables. Then, by using machine learning-based regression methods, we identify the primary meteorological variables for PV output power estimation. Finally, the paper concludes that the impact of using a lower-dimensional subspace of meteorological variables per location, as input for the regression methods, results in a similar estimation accuracy in the two case studies.

© 2020 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

## 1. Introduction

Due to the global concerns about climate change, renewable energy technologies are entering the energy production landscape rapidly. In recent years, there has been a sharp increase in the deployment of photovoltaic (PV) systems as a source of power generation in both standalone and grid-connected residential and large-scale systems [1]. It is expected that the cumulative installed capacity of PV could reach 22% of global electricity generation in 2050 [2]. This level of PV will support the transition into a more sustainable energy system and deliver substantial benefits in terms of security of energy supply and socio-economic development [3].

The output power of PV systems depends mostly on the global irradiance arriving on the plane of the PV array (POA). The

variability associated to the PV output power is caused by fluctuations in POA irradiation and can be divided into a deterministic and a stochastic component. The deterministic part is explained by the movement between the Sun and Earth, and follows a diurnal cycle. More significant and unexpected fluctuations in PV output power are caused by the stochastic behavior of the atmosphere. This uncertainty is due to multiple meteorological variables, such as temperature, humidity level, visibility, air pressure, wind speed and cloud cover [4,5].

The variable output power results in voltage and frequency fluctuations, which poses challenges on grid operation especially at high PV penetration rates. In order to cope with the power variability and maintain reliable grid operation, large amounts of costly storage facilities, balancing energy and/or frequency reserves are required [6,7]. Accurate PV output power forecasts have the potential to lower reserve capacity and maintain reliable grid operation as power production and consumption can be scheduled accordingly. This can potentially decrease the integration costs associated with high PV penetration. As a result, the field of PV output power estimation and forecasting has received increasing

\* Corresponding author.

E-mail addresses: [t.a.alskaif@uu.nl](mailto:t.a.alskaif@uu.nl) (T. AlSkaif), [soumyabrata.dev@adaptcentre.ie](mailto:soumyabrata.dev@adaptcentre.ie) (S. Dev), [l.r.visser@uu.nl](mailto:l.r.visser@uu.nl) (L. Visser), [murhaf.hossari@adaptcentre.ie](mailto:murhaf.hossari@adaptcentre.ie) (M. Hossari), [W.G.J.H.M.vanSark@uu.nl](mailto:W.G.J.H.M.vanSark@uu.nl) (W. van Sark).

<sup>1</sup> Authors contributed equally.

attention among researchers over the past decade. Moreover, thanks to the availability of reliable, high resolution and real-time measurement and processing units, significant improvements have been made in the field of solar power estimation and forecasting [4,5,8,9].

PV output power is either estimated directly or indirectly. Indirect techniques measure or predict the global horizontal irradiation (GHI) and then calculate the PV output power by means of a physical conversion model. Direct methods estimate or forecast the PV output power directly from meteorological variables and/or historical PV power measurements. Different methods of PV power forecasting can be generally divided into four classes, i.e. machine learning (ML), cloud imagery, physical and hybrid methods [4,5,10]. ML-based methods cover both statistical and computational intelligence methods. Cloud imagery methods consider either satellite images or all-sky imaging that in principal estimate GHI by determining cloud cover indices and cloud motion vectors. Physical methods include numerical weather prediction (NWP) models that predict irradiation by describing the development of the atmospheric state in time. Finally, hybrid models try to capture the strengths of different methods by combining two or more models [10,11]. The choice for forecasting methods highly depends on the desired forecasting horizon and spatial resolution.

Forecasting methods integrate various meteorological variables as exogenous input in order to achieve more accurate forecasts since these variables are considered as a source of uncertainty in PV output power [4,5,12]. PV output power forecasting using exogenous data has been addressed using different methods of forecast in Refs. [8,13,14]. The work in Ref. [13] uses a forecasting model based on satellite images and a support vector machines (SVM) learning scheme. A PV output power forecasting in a network of neighboring PV systems is proposed in Ref. [8]. The work is based on the cross-correlation time lag between clear-sky index of those PV systems that are influenced by the same cloud. In Ref. [15], an ANN forecasting method is proposed based on weather classifications considering the meteorological variables relative humidity, wind speed and air temperature in addition to solar irradiation. Another ML-based PV power forecasting approach [16] takes the meteorological variables temperature, cloud cover, wind speed, wind direction, relative humidity, air pressure and visibility into consideration. The work in Ref. [17] present a forecasting approach that is merely based on cloud cover data and PV power measurements. In Refs. [5,9,18], various other recently proposed exogenous forecasting methods can be found that consider one or several meteorological variables as input data. A particular exogenous model is presented in Ref. [19]. The proposed method trains a submodel for different types of daily weather classifications, i.e. a sunny day. Next, depending on the forecasted weather classification a certain submodel is called.

Despite their relevance, most of the exogenous forecasting methods of PV output power consider only few and different meteorological variables as data input. Besides, these studies focus mainly on the forecasting models and their final performance. Subsequently, in literature less attention is paid to the interdependence between different meteorological variables and their individual importance in the results of different methods of estimation and forecasting. An analysis of multiple meteorological variables was proposed in Ref. [20], but for the purpose of rainfall detection.

This paper aims to contribute to this research area and presents a systematic analysis of different meteorological variables that affect PV output power estimation. Based on two sets of 3 year-long data holding several meteorological values and PV output power measurements, we point out the variables that are most significant

to consider when estimating the PV output power (i.e., resulting in a lower-dimensional subspace of input meteorological variables), while we explicitly exclude solar irradiance as we are interested in how other meteorological variables that are less obvious affect PV output power. The PV output power estimation performance is then evaluated using different well-established ML-based regression methods and considering different sets of input variables. To increase the reliability of the results and capture potential deviations, the analysis and assessment of the estimation performance are performed using data sets in two regions with different climates, namely Austin, Texas, the U.S. and Utrecht, the Netherlands. These data sets are used to analyze the interdependence of different meteorological variables, the importance of the variables when estimating the PV output power, and the impact of the climate on the results.

Our study provides new insights on the interdependence and importance of a wide set of meteorological variables for PV output power estimation, and present a comprehensive comparison between various ML-based regression methods in terms of estimation errors. The methodology presented provide guidance in selecting the most important weather variables for estimating or forecasting PV output power at any location. In addition, the analysis gives explanations on why the estimation performance of some models deviate in different regions even when using the same input meteorological variables.

The structure of the paper is organized as follows.<sup>2</sup> Section 2 provides a description of the system and input data used in this study and shows how the relation between meteorological variables deviate per climate. The ML-based regression and variables importance methods are described in Section 3. In Section 4, we evaluate the estimation performance and analyze the impact of different meteorological variables on PV output power estimation. Finally, we conclude the paper and provide pointers for future work in Section 5.

## 2. System and data description

### 2.1. Input data sets

This study considers two case studies, namely Austin, Texas, the U.S and Utrecht, the Netherlands. For each location an input data set of meteorological variables (i.e., 3 year-long) is collected from nearby weather stations. Both data sets comprise a wide set of meteorological variables, namely ambient temperature (T), dew point temperature (DP), humidity level (RH), visibility (V), air pressure (P), wind speed (WS), cloud cover (CC), wind bearing (WB) and precipitation (R). Most of these meteorological variables are self-explanatory. Amongst them, WB is the direction of the flow of winds, measured in degrees. The cloud cover records in the two case studies are provided in oktas. Despite the fact that solar irradiance has a high correlation with the PV output power, this variable is not considered as an input in the analysis because our aim is to assess how other variables, with less obvious relationship, affect PV output power estimation. Moreover, this variable is available only for the Utrecht case study and not for Austin. Because of the strong relation between the diurnal solar cycle and the PV output power and the deterministic nature of this variable [5], the hour of the day (HoD) is considered as an additional input variable. Since the HoD has a cyclic nature, this variable consists of a x and y component that reflect the sine and cosine components of the HoD:

<sup>2</sup> The source code of all simulations in this paper will be provided as supporting information.

$$\text{HoD}_x = \cos \frac{2\pi * \text{HoD}}{24}, \quad (1)$$

$$\text{HoD}_y = \sin \frac{2\pi * \text{HoD}}{24}. \quad (2)$$

Another two data sets consisting of PV output power measurements (i.e., 3 year-long) are collected from multiple sites (i.e., households) in Austin and Utrecht. We performed an extensive quality assessment regarding the completeness of the data sets of each PV system. Further, any output power measurement at certain timestamps that was determined to be impossible has been removed (e.g., an hourly average production above the installed capacity or a negative production value during times of daylight). Next, more elaboration on the considered case studies is provided in Section 2.2.

## 2.2. Case studies

The data used in the first case study is collected from the Pecan Street Dataport [21], as part of the Pecan Street Demonstration, a smart grid research project located in Austin, Texas, US [22,23]. From the Dataport, we use 3 years of PV output power data from January 2014 until December 2016 of 24 households with rooftop PV systems (i.e., prosumers). This data is available in the Pecan Street Dataport in an hourly resolution. The installed capacity of these PV systems vary between 2.9 and 8.8 kWp and all PV panels are orientated to the south. We use the meteorological variables data (i.e., available in the same Dataport) for the same location and with the same data resolution as the PV data. The households are located in Austin, Texas, US, with coordinates 30.2672° N, 97.7431° W. According to the Köppen-Geiger climate classification, Austin has a humid-subtropical climate (i.e., Cfa) [24].

In the second case study, 3 years of hourly meteorological variables data from February 2014 until January 2017 are used to generate the results. The data are measured by a weather station at De Bilt, Utrecht, the Netherlands with coordinates 52.0907° N, 5.1214° E. This weather station belongs to the Royal Netherlands Meteorological Institute (KNMI) and the data is made available online via their website [25]. Measurements of rooftop PV output power of 10 households located near the weather station are used. These measurements are available in 1-min resolution and averaged for every hour of the assessment period in order to provide a fair comparison with the other case study. The size and orientation of these PV systems respectively ranges from 0.5 to 3.0 kWp and 175–185° (i.e., south orientated). For more information about the rooftop PV systems, please refer to Refs. [8,12]. Other than Austin, the climate in Utrecht is classified as Oceanic (i.e., Cfb) [24].

In both case studies the PV output power is measured directly, meaning that it represents the PV modules' generation. Those data sets are used as input to the regression methods in order to determine the most important meteorological variables. Moreover, nocturnal timestamps have been removed from all data sets (i.e., when PV output power is 0), such that these night-time values are not considered in training nor evaluation. In both case studies, the first two years of data are used for training, the third year is used to test and compare the performance of the methods. Since the size of the PV systems vary per site, the output power of each system is normalized according to the installed capacity before the regression methods are trained. Consequently, all values and errors are independent of the system size and can be compared directly.

In order to show how the relation between meteorological variables deviate per climate, in the next sections we assess their

interdependence and perform a principle component analysis (PCA).

## 2.3. Interdependence of meteorological variables

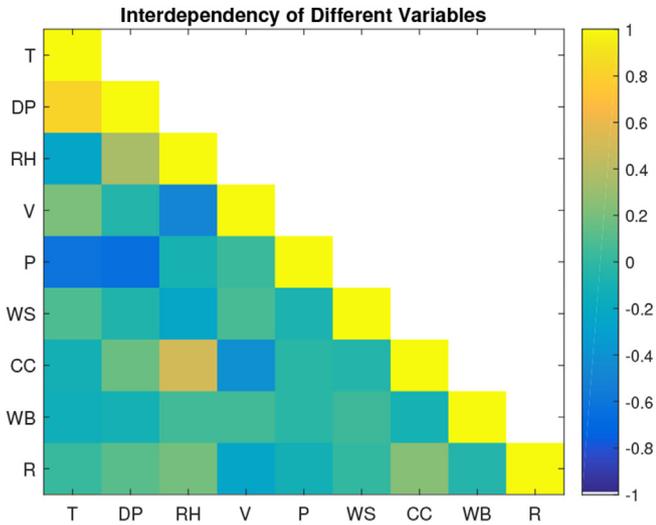
This section analyses the interdependence of  $n = 9$  meteorological variables in each site, using the 3 years of meteorological data described in Section 2.1, and highlights how this is different in the two case studies. We consider different vectors of meteorological variables which are indicated by  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ , where  $n$  is the number of considered meteorological variables. Each vector is of length  $m$  which indicates the weather recordings for all timestamps in the assessment period. If we stack all the column vectors, we get the variable matrix  $\mathbf{X}$ , of dimension  $m \times n$ , as:

$$\mathbf{X} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n]. \quad (3)$$

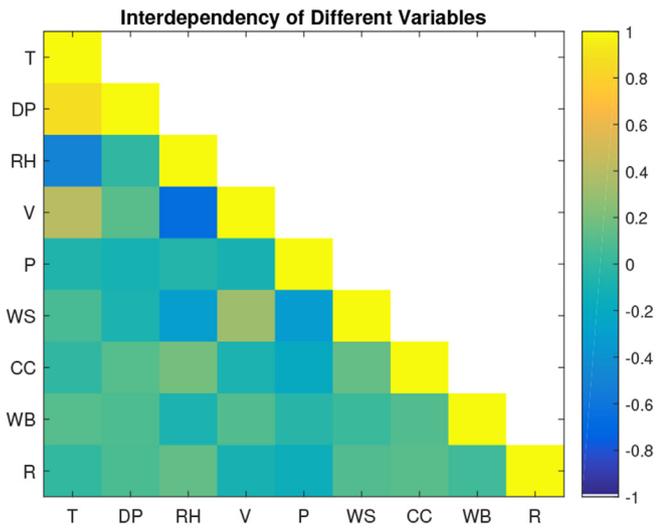
The cross-correlation value among the various meteorological variables is computed by calculating the correlation coefficients of the variable matrix  $\mathbf{X}$ . The correlation coefficient between a pair of meteorological variables vectors indicates the degree of correlation between them (i.e., a measure of their linear dependence) [26].

The resulted interdependence of the considered meteorological variables in the 3 year period is illustrated in Fig. 1. The off-diagonal elements show the degree of dependency for each pair of variables. It can be observed from the figure that in the two considered case studies, the temperature (T) has a strong positive correlation with the dew point temperature (DP) indicating that when T increases, the DP also increases (i.e., the correlation of T (x-axis in Fig. 1) with DP (y-axis) is 0.81 in Austin and 0.87 in Utrecht). The visibility (V) has a strong negative correlation with the relative humidity (RH) (i.e.,  $-0.48$  in Austin and  $-0.67$  in Utrecht) indicating that when the horizontal view during observation is high, the RH is low, and vice versa. Furthermore, V is positively correlated with T (e.g., during sunny days in spring and summer seasons) and somehow with wind speed (WS) (i.e., when WS is high, V is usually high). However, Fig. 1 (a) and (b) show that this positive correlation is significantly higher for Utrecht than Austin (e.g., correlation of T with V is 0.21 in Austin and 0.41 in Utrecht). Similarly, cloud cover (CC) has a positive correlation with RH (i.e., indicating that in an overcast condition, the RH has a high value). This correlation is higher, however, in Austin (0.5) than in Utrecht (0.21). In the Netherlands, it is likely that the humidity is high when it is cloudy but this can also happen during less cloudy days due to the relative large inland water surface area (i.e., a fifth of the total surface area consists of water in the Netherlands). Moreover, the RH is found to be more stable in Utrecht throughout the year including the summer when compared to Austin. In addition, RH is in both locations negatively correlated with temperature (i.e., as T increases, RH drops).

Besides the overlapping correlation of several meteorological variables in both locations, some (anti-) correlations are only found at a single location. In Austin, V and CC are negatively correlated. In addition, both T and DP are strongly negatively correlated with air pressure (P). This makes sense because pressure decreases with an increasing temperature, especially in a humid-subtropical climate as found in Austin. In Utrecht, wind speed is negatively correlated with both P and RH (i.e., when WS is high P and RH are usually low, note that P and RH are not correlated). Understanding these (anti-) correlations and dependencies help in explaining the performance dependency of the regression models on the input meteorological variables considered. Moreover, it provides guidance in selecting and ranking the meteorological variables for classification and regression tasks in the field of solar energy analytics and forecasting.



(a) Austin.



(b) Utrecht.

**Fig. 1.** Interdependence of the various meteorological variables amongst each other. Temperature (T), dew point temperature (DP), humidity level (RH), visibility (V), air pressure (P), wind speed (WS), cloud cover (CC), wind bearing (WB) and precipitation (R) (best viewed in color). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

2.4. Principal component analysis (PCA)

In this work, the PCA method is used to further analyze and better understand the interdependence between the considered  $n$  meteorological variables in each case study. In general, PCA is an unsupervised linear transformation technique that is prominently used for feature extraction and dimension reduction [27]. Using PCA, it is possible to transform the input meteorological variables data onto a new feature space that maintains the most relevant information. This can be done by finding the directions of maximum variance in the input data sets and project it onto a new subspace with equal or fewer dimensions than the original one.

To find the principal components, first the variable matrix  $X$  is standardized across each of the meteorological variables, with respect to the mean and standard deviation of each vector. Then the covariance matrix of the standardized variable matrix is constructed, which stores the pairwise covariances between the different meteorological variables. A positive covariance between two variables indicates that the variables increase or decrease together, whereas a negative value indicates that the features vary in opposite directions. After that, an eigenvectors and eigenvalues decomposition of the covariance matrix is performed and the resulted eigenvectors define the new orthogonal components, also called “the principal components”, whereas the corresponding eigenvalues will define their magnitude. After sorting the eigenvalues by decreasing order and ranking the corresponding eigenvectors, the first principal component will have the largest possible variance (i.e., information), followed by the second principal component with second largest variance and so on. It is worth noting that the resulting principal components are mutually uncorrelated to each other even if the input variables are correlated due to the orthogonality of the decomposed eigenvectors [27].

Fig. 2 illustrates the percentage of variance captured by each of the resulting principal components using the  $n = 9$  input meteorological variables we consider in this study. The plots indicate that in both locations the first principal component alone accounts for approximately 27% of the variance in the input dataset. This indicates that there is a high degree of correlation amongst the input variables. Moreover, it can be observed from Fig. 2, that cumulatively, around 86.3% and 80% of the variance in the data sets is captured by the first 5 principal components in Austin and in Utrecht, respectively. This means that the PCA enables to reduce the amount of input variables. However, the whole set of input variables is needed to calculate the principal components in the first place.

The biplot representation of the input meteorological variables is depicted in Fig. 3. The figure illustrates the importance in terms of the contribution of the input variables into the first two primary principal components and confirms the correlation results in Fig. 1. It can be noticed that the data is more spread along the x-axis (i.e., principal component 1) than the y-axis (i.e., principal component 2), which is consistent with the results of Fig. 2 and the percentage of variance captured by each of the principal components (i.e., highest for principal component 1 in both case studies).

The biplots also show that the meteorological variables are captured differently in the principal components that are generated in both case studies as a result of the local climates. For instance, DP has a high contribution to principal component 1 and 2 in Utrecht, whereas this is only limited to principal component 1 in Austin. In Utrecht CC and R are only captured in principal component 2 while these variables contribute to both principal components in Austin.

Additionally, the biplots confirm the results we found in Fig. 1. Moreover, Fig. 3 shows that T and DP are positively correlated with each other, whereas P and DP are highly correlated with a negative value. RH is at both locations strongly negatively correlated with V. In addition, RH and CC are strongly positively correlated with each other in case of Utrecht and significantly less correlated in Austin.

3. Methods

3.1. Regression methods

In this section, several ML-based regression methods are presented. These methods will be used to estimate the PV output power  $p$  using the considered meteorological variables (see Section 2). There are  $n$  meteorological variables considered as input to the estimation methods. The response vector  $p$  and each vector of the

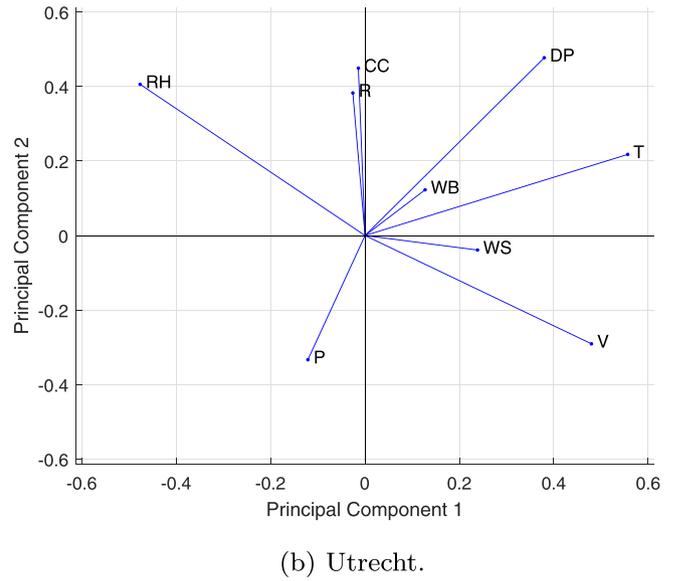
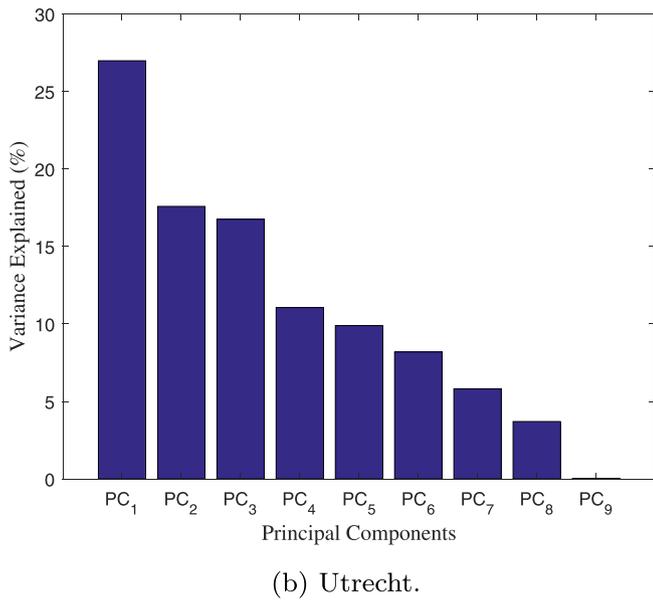
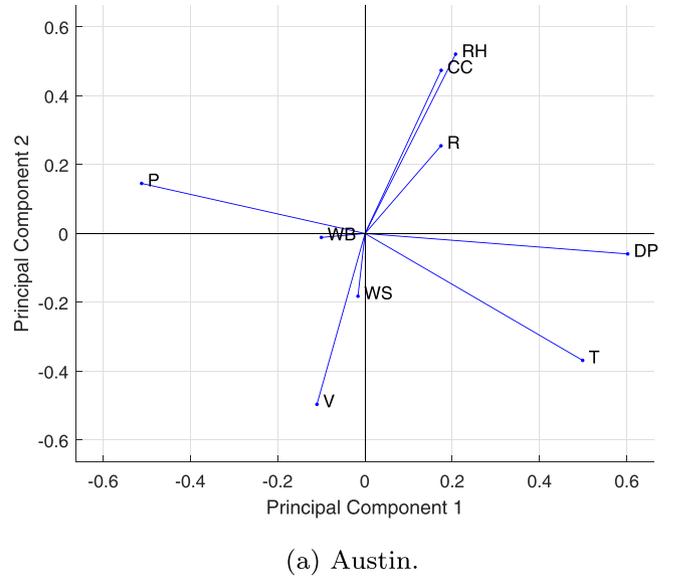
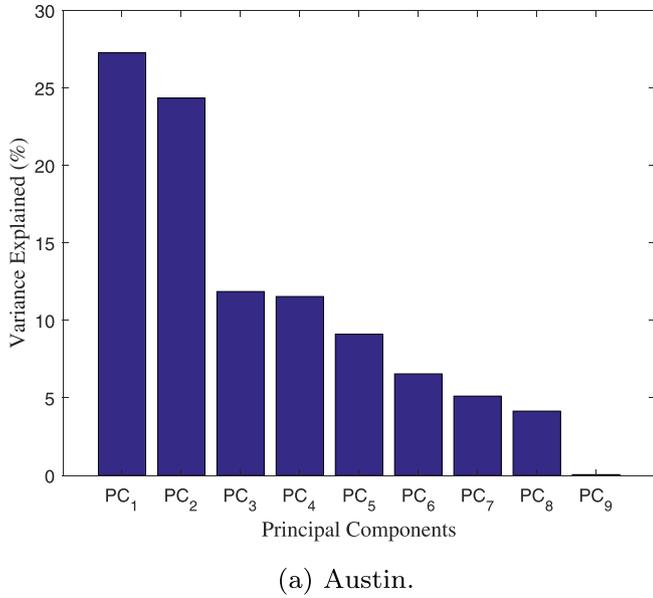


Fig. 2. Distribution of the captured variance across the several principal components.

meteorological variables are of the dimension  $m \times 1$ , where  $m$  indicates the number of observations in the considered assessment period. Without the loss of generality, unless otherwise stated, we will drop the index  $m$  from the subsequent discussion.

### 3.1.1. Multivariate linear regression (MLR)

Multivariate linear regression is one of the fundamental and most widely used regression methods. The MLR model estimates the PV output power,  $\mathbf{p}^{MLR}$ , as:

$$\mathbf{p}^{MLR} = \beta_0 + \beta_1 \mathbf{v}_1 + \beta_2 \mathbf{v}_2 + \dots + \dots + \beta_n \mathbf{v}_n, \quad (4)$$

where the  $\beta$ s are the regression co-efficients.

### 3.1.2. LASSO regression (LR)

Least absolute shrinkage and selection operator (LASSO) is one

Fig. 3. Biplot representation of the meteorological variables in the first two principal components.

of the common methods to regress the PV output power. It estimates the PV output power by penalizing the regression coefficients (i.e.,  $\beta$ s). Such penalty is useful to shrink the input variables that are not very important in estimating the output variable. If the applied penalty is too large, the estimation is shrunk to zero. In this sense, LASSO is considered as a continuous feature selection method [28]. This is useful in case some of the input meteorological variables are correlated to each other. The regressed PV output power using LASSO is calculated by solving the following optimization problem:

$$\text{minimize } (\mathbf{p} - \mathbf{v}\beta)^\top (\mathbf{p} - \mathbf{v}\beta), \text{ s.t. } \sum_1^n |\beta_i| \leq h, \quad (5)$$

where operator  $\cdot^\top$  denotes the vector transpose,  $\mathbf{v}$  is the training data of all meteorological variables and  $h$  is a user-defined

threshold.

3.1.3. SVM regression

SVM regression, also known as SVR [29], has been shown to obtain a good performance in the area of PV output power forecasting and estimation since it keeps low complexity and a good fitting of data [30,31]. In this study, we use the non-parametric SVM regression to estimate the PV output power. The linear SVM (L-SVM) regression method attempts to estimate the PV output power using the following optimization problem:

$$\text{minimize } \frac{1}{2} \|w\|^2, \text{ s.t. } \mathbf{p} - \langle w, \mathbf{v} \rangle + b \leq \epsilon, \tag{6}$$

where  $w$  is the weight, and the estimated PV power is  $\langle w, \mathbf{v} \rangle + b$ . The parameter  $\epsilon$  serves as a threshold.

The kernel, or non-linear, SVM (K-SVM) regression has the same formulation, with the dot product  $\langle w, \mathbf{v} \rangle$  replaced by a kernel function.

3.1.4. Random forests (RF)

Random forests is an ensemble-based regression method that consists of an ensemble of decision/regression trees, whose results show the mean prediction of individual trees. In RF a number of decision trees is generated with a set of  $n$  layers for each tree. At each layer in every tree, there are  $2^n$  decision nodes with  $n = 0$  at the first layer. Every decision node has its own characteristic variable conditions and based on these conditions, the node will pass a true or false to the node in the next layer ( $n + 1$ ). In the last layer (i.e., leaf layer), this will result in an estimation of the target value based on the average of all samples reaching that node. Every tree in the random forest is trained using a random subset of the training data. The same training data can be selected by different trees in the forest [32].

Random forests has recently been considered as a promising method in the area of PV output power forecast [9,16]. We use random forests in our work to fit the response vector  $\mathbf{p}$  using the training input variables vectors  $\mathbf{v}$ , and train 100 regression trees using the least squares boosting (LSBoost) algorithm [32]. The ensemble attempts to fit a new learner, at every iteration, by computing the difference between observed response value, and the accumulated prediction of all learners developed previously.

3.2. Variables importance

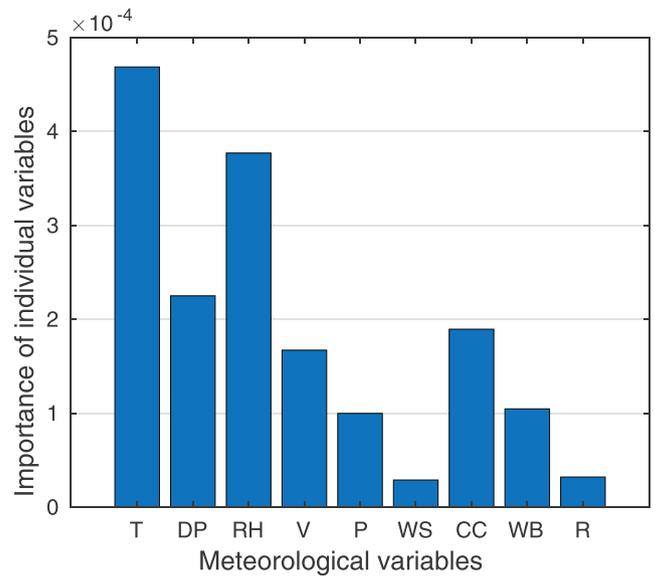
This section provides a method to analyze the importance of each vector of input meteorological variables when estimating the PV output power. To do that, a regression ensemble model is trained using the response vector  $\mathbf{p}$ , and the predictor variables vectors  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_9$  in  $\mathbf{X}$ . The MATLAB function `predictor-Importance` is then used to estimate the importance of each of these predictor vectors in regressing the PV output power  $\mathbf{p}$ . It estimates the importance of a predictor variable vector for the regression ensemble by accumulating the estimates over all the weak learners in the ensemble. The output of the function has one element for each input predictor in the data used to train the ensemble – a higher value for this element indicates that the particular variable has more importance in regressing the PV output power.

4. Results and discussions

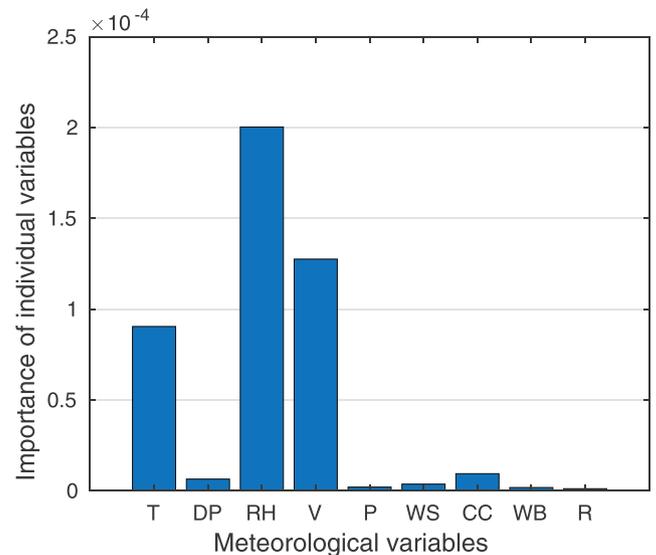
4.1. Importance of meteorological variables

The same 9 meteorological variables are used to perform the

importance analysis for the case studies in Austin and Utrecht over a period of 3 years. The least squares boosting (LSBoost) algorithm is used as a regression ensemble model for the `predictor-Importance` MATLAB function. Fig. 4 shows the importance of all input predictor variables  $\mathbf{v}_i$  in PV output power estimation. The figures show some distinctive differences between the studied locations. In general, in Austin more variables are found to be of importance in estimating the PV output power compared to Utrecht. In addition, the order of the most important variables does not overlap for both locations (e.g., T and RH are found to be the most important variable for Austin and Utrecht, respectively). Subsequently, DP is found to be the fifth important variable in Utrecht, whereas it is the third in Austin. The limited importance of DP in Utrecht could be explained by the strong correlation observed between DP with T (see Fig. 1). Similarly, RH is found to have a strong negative correlation with T in Utrecht, which may explain

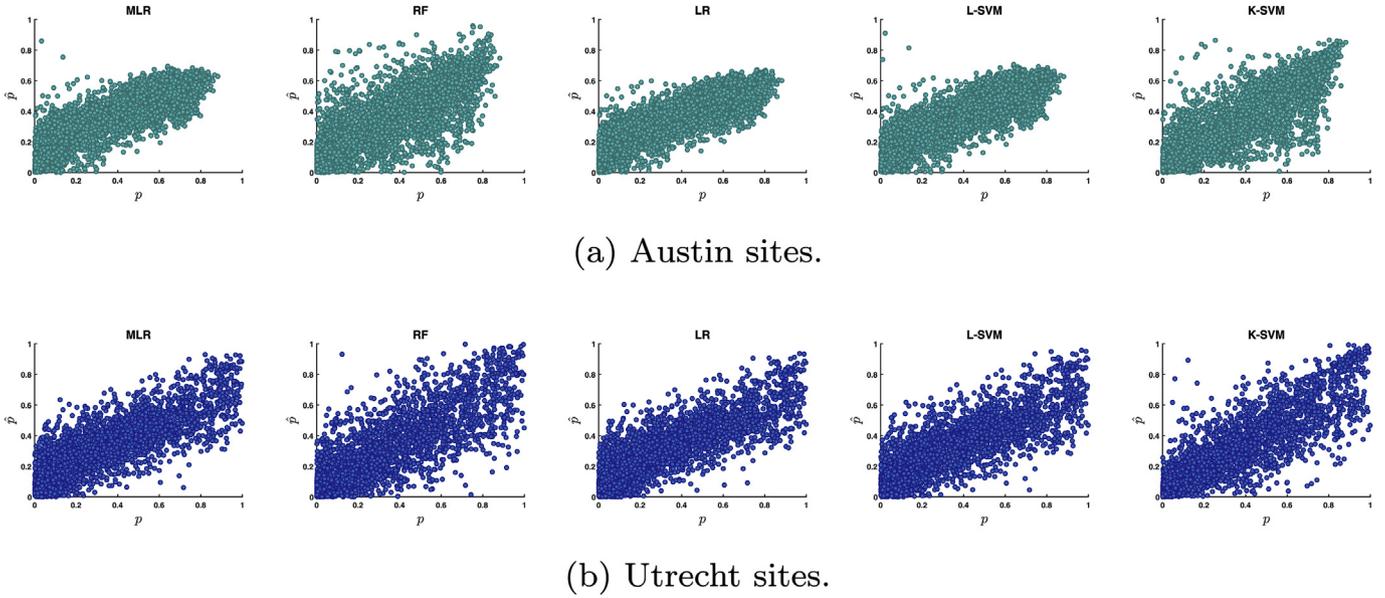


(a) Austin sites.



(b) Utrecht sites.

Fig. 4. Importance of the meteorological variables in estimating the PV output power.



**Fig. 5.** Scatter plots between the measured ( $p$ ) and estimated ( $\hat{p}$ ) PV output power values (for  $n = 9$ ) for an arbitrary PV system in Austin (a) and Utrecht (b). The regression methods are: multivariate linear regression (MLR), random forests (RF), LASSO regression (LR), linear support vector machine (L-SVM) and kernel support vector machine (K-SVM).

the lower importance of T in that location. However, also some similarities can be observed. In both locations, although in a different order, RH, T and V belong to the most important variables. CC scores the fourth variable for estimating PV output power in both case studies. However, CC is almost not important for Utrecht and shows a different behavior in Austin. This could be related to the cloudiness measurements available in oktas which could be a better proxy for estimating PV output power in Austin than in Utrecht.

It is important to note that Fig. 4 explains which predictors are the best indicators for estimating PV output power rather than what variables affect the PV output power generation. For instance, CC could affect PV generation more than T, however, T proves to be more correlated or more important for estimating PV output power. The figure also shows that the importance and ranking of the variables depend on the climate of the considered area of study.

#### 4.2. Estimation performance evaluation

In this section, we evaluate the different benchmarking ML-based regression methods mentioned in Section 3. In addition, we compare the performance of each method when using the original dimension of meteorological variables (i.e.,  $n = 9$ ) with the performance when using a lower-dimensional subspace of meteorological variable. The input of the models consists of 3 years data of meteorological variables and PV output power of multiple PV systems in Utrecht and Austin (see Section 2). The meteorological variables considered in the reduced subspace are the  $n = 4$  variables that are found to be most important in describing the PV output power, which are in order T, RH, DP and CC for Austin, and RH, V, T and CC for Utrecht (see Section 4.1). In addition, since Fig. 4 shows a different magnitude in the importance of the meteorological variables between the two case studies, additional experiments for regressing the PV output power are carried out when considering only the top  $n = 3$  variables. The performance evaluation and comparison of the regression methods are achieved using different performance metrics, namely mean absolute error (MAE), root mean squared error (RMSE), mean squared log error (MSLE) and mean bias error (MBE). These performance metrics are

respectively defined as follows:

$$\text{MAE}(p, \hat{p}) = \frac{1}{m_{\text{samples}}} \sum_{t=0}^{m_{\text{samples}}-1} |p_t - \hat{p}_t|, \quad (7)$$

$$\text{RMSE}(p, \hat{p}) = \sqrt{\frac{1}{m_{\text{samples}}} \sum_{t=0}^{m_{\text{samples}}-1} (p_t - \hat{p}_t)^2}, \quad (8)$$

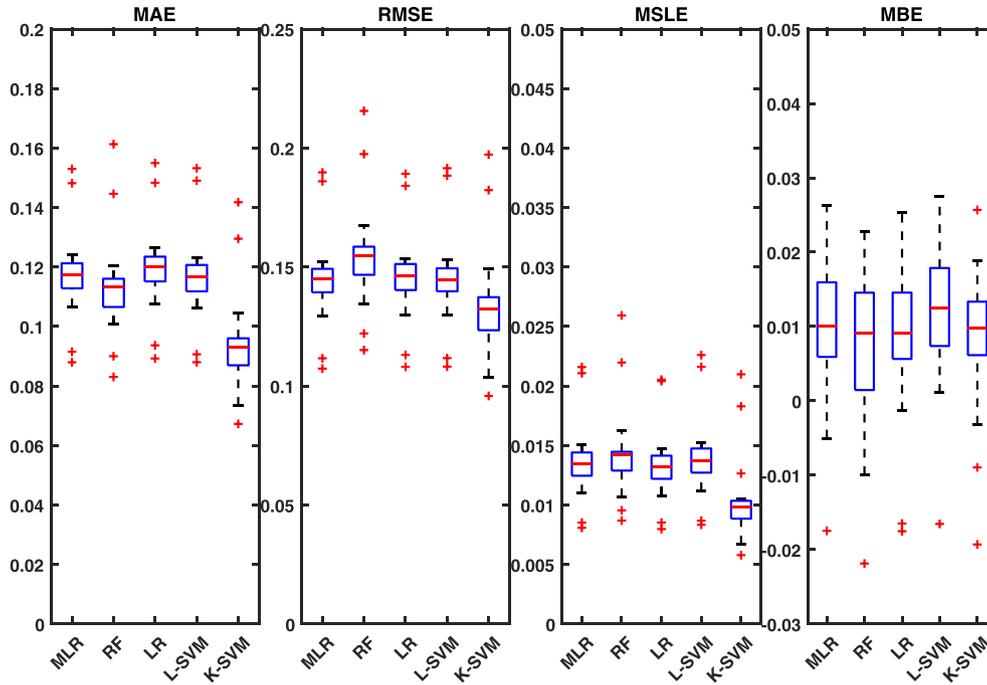
$$\text{MSLE}(p, \hat{p}) = \frac{1}{m_{\text{samples}}} \sum_{t=0}^{m_{\text{samples}}-1} (\log_e(1 + p_t) - \log_e(1 + \hat{p}_t))^2, \quad (9)$$

$$\text{MBE}(p, \hat{p}) = \frac{1}{m_{\text{samples}}} \sum_{t=0}^{m_{\text{samples}}-1} (p_t - \hat{p}_t). \quad (10)$$

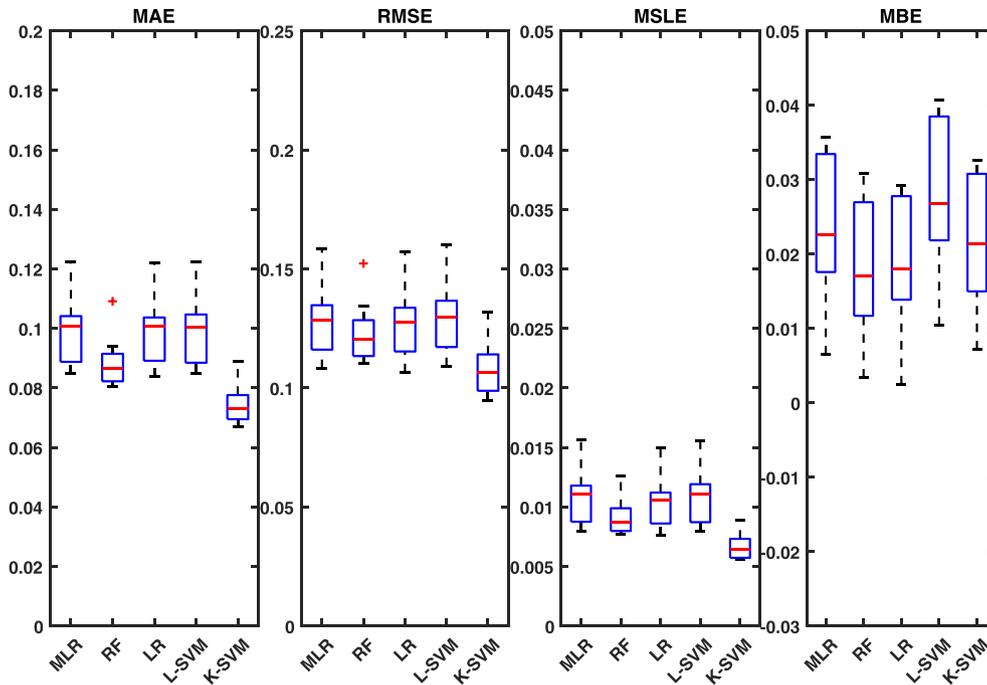
As described in Section 4.1, the first two years of data describing the meteorological variables and the PV output power of each PV system are used as a training set for all the regression methods. The regression coefficients of all methods are learned from the output power of each PV system individually. The trained models are then used for estimating the PV output power for the third year of the initial data sets. The results are then evaluated by the error metrics presented above.

Fig. 5 shows scatter plots holding the measured  $p$  and estimated  $\hat{p}$  PV output power values for an arbitrary PV system in the two case studies and using the five different regression methods described in Section 3. The whole set of predictor variables (i.e.,  $n = 9$ ) are used to estimate the PV output power. These scatter plots show how the estimation by each method is distributed along the PV output power measurement. From the figure it can be observed that the linear regression methods are slightly skewed in a way that low PV output power values are often overestimated and higher PV output power values are repeatedly underestimated. The nonlinear K-SVM and RF methods appear to be slightly less vulnerable to this.

Fig. 6 presents the summarized results of the different



(a) Austin sites.



(b) Utrecht sites.

**Fig. 6.** Performance evaluation of the benchmarking regression methods on 24 distinct households in Austin and 10 in Utrecht when considering all meteorological variables ( $n = 9$ ). All experiments are performed individually for all households. In each box plot, the top and bottom blue edges indicate the 75th and 25th percentiles, respectively. The central red mark represents the median and the outliers are indicated by the red '+' symbols. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

regression methods for all PV systems in both case studies. Consequently, the box plots holds the error value of each PV system considered in the case study of interest. Similarly, the  $n = 9$  meteorological variables are used to estimate the PV output power

in this experiment. From Fig. 6 a general trend can be observed regarding the performance of the different estimation methods. Firstly, in both case studies the K-SVM regression method shows a better estimation performance in terms of the MAE, RMSE and

MSLE. This means that K-SVM outperforms the other methods in capturing the nonlinear relationship between the input meteorological variables and the response PV output power. Secondly, the MLR, LR and L-SVM regression methods are found to achieve very similar results at each location. Moreover, these comparable results are likely the outcome of the fact that all those three methods describe a linear relationship between the predictor and response variables. Finally, the RF method is the second best performing method in the case of Utrecht. This can be explained as the RF method is capable to some extent of capturing the nonlinear relationship between the input and output variables. In terms of the MAE, this also holds for Austin. However, it is remarkable that when we consider the RMSE, we find that all methods outperform RF in the case of Austin. A lower observed MAE and a higher RMSE implies that RF performs better in general, but at the same time is coupled with more extreme outliers. Consequently, the PV output power estimate by the RF method proves to be more often significantly wrong. This can also be observed in Fig. 5, where the scatter plot of RF in Austin is more cloudy. The box plots in Fig. 6 also present the MBE observed per method. A positive bias is observed for all methods in both Austin and Utrecht, which implies that all models underestimate the PV output power. This underestimation is more significant for Utrecht than Austin. In the latter, a number of methods also overestimate the total output power of a few PV systems which are indicated as the 25th percentile and/or some outliers.

#### 4.3. Reduced subspace estimation performance evaluation

The estimation performance of the different regression methods when reducing the space of input meteorological variables is presented in Table 1. The table shows the average increase/decrease in estimation error of the different performance metrics when a lower-dimension subspace (i.e.,  $n = 4$  and  $n = 3$ ) of predictor variables is used compared to the full space (i.e.,  $n = 9$ ). In the table, a negative sign (–) indicates a performance improvement.

From the table it can be concluded that the two case studies, Austin and Utrecht, are very differently affected by reducing the space of input meteorological variables for the regression methods. This impact is more significant in Austin than Utrecht. The results in the table for Austin indicate that the regression methods MLR, RF and L-SVM perform slightly better when  $n = 4$  meteorological variables are considered instead of  $n = 9$ . This increased uncertainty in the PV output power estimation when considering a larger number of input variables can be caused by two main reasons. First, as Fig. 4 indicates, the importance of many of the input variables is limited, such that the predictors hold limited information for

estimating the PV output power. Therefore, considering those variables cloud the estimation and add to the estimation uncertainty. Another explanation could be that by selecting only  $n = 4$  variables, some variables are excluded which have a high (anti) correlation with some of the selected variables. For example, in Fig. 1 the variable V, which is excluded when  $n = 4$ , showed to be highly negatively correlated with the selected variables CC and RH. By excluding the correlated variable, the regression methods are less prone to issues that may occur as a result of the effect multicollinearity has on estimating the model coefficients. Similar to MLR, RF and L-SVM, the performance of the K-SVM method also improves in terms of the MAE when  $n = 4$  variables are selected. However, the performance of this method worsens in terms of the RMSE and MSLE. This implies that as the input variables reduced to  $n = 4$ , the estimated values hold more outliers and higher relative errors. In contrast to the other methods, the performance of LR in terms of the MAE, RMSE and MSLE lessens as less information is available from the input variables considered. This indicates that LR is able to deal with the issues discussed above and extract some additional information of the correlated or less important variables. Finally, in Austin a significant poorer performance for all methods is found when  $n = 3$  predictor variables are considered. Therefore, it is found that excluding CC highly affects the performance of all regression methods.

In Utrecht, it is noticed that except for RF, the performance of all methods is getting slightly worse when the top  $n = 4$  meteorological variables are considered. The improved importance of RF may be explained as the additional variables, that were found to be of very limited importance in Fig. 1, cloud the performance of the regression method. Due to the random features considered while training the RF method, this could affect RF more than the other regression methods. All methods are found to perform worse when considering the top 3 predictors only. Remarkably, removing the CC as an input variable affects the performance of K-SVM significantly more compared to the other methods. From Fig. 4 it can be observed that CC is significantly less important than the top 3 variables. This could explain the limited impact of removing the variable in estimating the PV output power in case of MLR, RF, LR and L-SVM. Moreover, from the results it can be concluded that K-SVM is able to capture the additional information that CC holds. Although less significant, the same trend can be observed in Austin. The capability of K-SVM to extract this additional information translates into a higher model performance that we observed in Fig. 6. Finally, although not significant from the table it can be observed that the performance of LR improves in terms of the MSLE. This implies that there are less relative large estimation errors when less predictor variables are considered. The same can be

**Table 1**  
Average increase/decrease in estimation error of the different performance metrics for the different regression methods when using the reduced space (i.e., top 4 and 3 variables) compared to the full space (i.e., 9 variables). The most important meteorological variables were found to be: RH, T and V and WB for Austin, and RH, V, T and CC for Utrecht (see Section 4.1), ordered from most important to less important variables. The decrease in estimation error is represented by a negative sign (–). All experiments are performed individually for each site in the two case studies before calculating the average.

	Methods	MAE		RMSE		MSLE	
		Top 4 variables	Top 3 variables	Top 4 variables	Top 3 variables	Top 4 variables	Top 3 variables
Austin sites	MLR	–0,2%	5,3%	–0,3%	8,4%	–1,0%	12,2%
	RF	–3,2%	2,5%	–1,7%	6,3%	–3,6%	10,8%
	LR	1,2%	6,4%	0,7%	8,4%	0,4%	12,6%
	L-SVM	–0,1%	5,7%	–0,3%	9,2%	–1,3%	13,6%
	K-SVM	–1,4%	8,3%	1,6%	16,1%	3,1%	34,0%
Utrecht sites	MLR	0,4%	1,0%	0,4%	1,3%	0,5%	–1,0%
	RF	–4,7%	1,5%	–1,9%	2,9%	–6,5%	3,6%
	LR	0,1%	1,0%	0,2%	1,1%	–0,4%	–0,6%
	L-SVM	0,6%	1,5%	0,4%	1,6%	1,1%	0,6%
	K-SVM	0,4%	6,2%	1,7%	6,9%	3,1%	16,0%

observed for the MLR model when  $n = 3$  input variables are taken into account.

As mentioned above and shown in Table 1, the average increase or decrease in estimation error when using a lower-dimension subspace, i.e.  $n = 4$  is relatively limited for all regression methods in the two case studies. The relative increase in errors becomes more significant for most regression methods when  $n = 3$  input variables are considered in the case of Austin as well as for the K-SVM method in Utrecht. This indicates that, depending on the regression method and location, a low number of meteorological variables recordings is enough to generate similar results without affecting the performance (i.e.,  $n = 4$  compared to the full space of  $n = 9$  variables). A similar effect may be expected in the case of solar power forecasting, where a lower-dimensional subspace could achieve a similar performance as achieved when considering multiple input variables. To increase the reliability of the results, the methods require testing under different conditions (e.g., additional climates, more PV systems and different PV systems installation settings). For instance, the performance could be different in other climates where other meteorological variables could be more important. As seen in Fig. 4, the selected variables in the lower-dimension subspace were different based on the climate of the area of study and therefore the performance of the regression methods. This might also explain why some exogenous forecasting models perform differently in different regions.

## 5. Conclusions and future work

This paper provided a systematic analysis of different input meteorological variables in the context of PV output power estimation. A complete 3 years of meteorological and PV output power data are used to establish the relation between the two data sets. Besides, the study provides methods to assess the interdependence of the meteorological variables and the importance of the variables in estimating the PV output power. Consequently, this study shows how the importance of variables and the estimation accuracy depends on the regression method and the climate zone.

In addition, we compared the performance of the regression methods when considering a lower-dimension subspace of predictor meteorological variables in two different case studies with different climate zones. The numerical evaluation showed that using a lower-dimension subspace of meteorological variables, as an input for the estimation methods, can result in a similar estimation accuracy. However, the results also show that both the most important input variables as well as the effect of selecting only the top variables on the performance highly depends on the location of interest. The proposed work may provide insights to test the performance of forecasting models based on lower-dimensional subspace instead of using all available input meteorological variables. The analysis and methods used in this study are generic and can also be used to perform similar analysis for other climate zones.

The work presented in this paper could be considered as a first step towards closing the research gap on which and how much meteorological variables are required for accurate PV output power estimation using ML techniques. Our future works will focus on using the established methodology and expanding the analysis to other locations of different climate zones. This would require worldwide records of meteorological data accounting to many different climates. In addition, we aim to assess and benchmark the estimation performance of PV output power by considering additional ML-based regression models.

## Author contribution

Tarek AlSkaif, Soumyabrata Dev, Lennard Visser:

Conceptualization, Methodology, Software, Data curation, Writing-Original draft preparation, Visualization, Investigation, Validation, Writing- Reviewing and Editing. Murhaf Hossari: Conceptualization, Visualization, Writing- Reviewing and Editing. Wilfried van Sark: Supervision, Visualization, Reviewing and 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.

## Acknowledgments

This work by the author and co-authors at Utrecht University is partly supported by the Joint Programming Initiative (JPI) Urban Europe project: “PARTicipatory platform for sustainable ENergy management (PARENT)”, the B-DER project 1621404 (funded by the Netherlands Enterprise Agency (RVO) within the Dutch Top-sector Energy framework), and eNErgy intrAneTs (NEAT) project (funded by the Netherlands Science Foundation (NWO)). The contribution of the co-authors at ADAPT Centre for Digital Content Technology is partially supported by the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

## References

- [1] A. Reinders, P. Verlinden, A. Freundlich, *Photovoltaic Solar Energy: from Fundamentals to Applications*, John Wiley & Sons, 2017.
- [2] IRENA, Global Energy Transformation, A Roadmap to 2050, Tech. rep., IRENA, Abu Dhabi, 2018. <http://www.irena.org/publications/2018/Apr/Global-Energy-Transition-A-Roadmap-to-2050>.
- [3] IEA, Technology Roadmap Solar Photovoltaic Energy, International Energy Agency (IEA).
- [4] M.Q. Raza, M. Nadarajah, C. Ekanayake, On recent advances in PV output power forecast, *Sol. Energy* 136 (2016) 125–144.
- [5] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F. Martinez-de Pison, F. Antonanzas-Torres, Review of photovoltaic power forecasting, *Sol. Energy* 136 (2016) 78–111.
- [6] T. AlSkaif, W. Schram, G. Litjens, W. van Sark, Smart charging of community storage units using Markov chains, in: IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe), IEEE, 2017, pp. 1–6, 2017.
- [7] T. Terlouw, T. AlSkaif, C. Bauer, W. van Sark, Multi-objective optimization of energy arbitrage in community energy storage systems using different battery technologies, *Appl. Energy* 239 (2019) 356–372.
- [8] B. Elsinga, W.G. van Sark, Short-term peer-to-peer solar forecasting in a network of photovoltaic systems, *Appl. Energy* 206 (2017) 1464–1483.
- [9] C. Voyant, G. Nottton, 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.
- [10] J. Kleissl, *Solar Energy Forecasting and Resource Assessment*, Academic Press, 2013.
- [11] M. Digne, M. David, P. Lauret, J. Boland, N. Schmutz, Review of solar irradiance forecasting methods and a proposition for small-scale insular grids, *Renew. Sustain. Energy Rev.* 27 (2013) 65–76.
- [12] L. Visser, T. AlSkaif, W. van Sark, Benchmark analysis of day-ahead solar power forecasting techniques using weather predictions, in: IEEE 46th Photovoltaic Specialist Conference (PVSC), IEEE, 2019, 2019.
- [13] H.S. Jang, K.Y. Bae, H.-S. Park, D.K. Sung, Solar power prediction based on satellite images and support vector machine, *IEEE Trans. Sustain. Energy* 7 (3) (2016) 1255–1263.
- [14] S. Dev, F.M. Savoy, Y.H. Lee, S. Winkler, Estimation of solar irradiance using ground-based whole sky imagers, in: Geoscience and Remote Sensing Symposium (IGARSS), IEEE International, IEEE, 2016, pp. 7236–7239, 2016.
- [15] C. Chen, S. Duan, T. Cai, B. Liu, Online 24-h solar power forecasting based on weather type classification using artificial neural network, *Sol. Energy* 85 (11) (2011) 2856–2870.
- [16] M.P. Almeida, O. Perpiñán, L. Narvarte, PV power forecast using a nonparametric PV model, *Sol. Energy* 115 (2015) 354–368.
- [17] D. Pepe, G. Bianchini, A. Vicino, Model estimation for solar generation forecasting using cloud cover data, *Sol. Energy* 157 (2017) 1032–1046.
- [18] Van der Meer, W. Dennis, Joakim Widén, M. Joakim, Review on probabilistic forecasting of photovoltaic power production and electricity consumption, *Renew. Sustain. Energy Rev.* 81 (2018) 1484–1512.
- [19] H.-T. Yang, C.-M. Huang, Y.-C. Huang, Y.-S. Pai, et al., A weather-based hybrid

- method for 1-day ahead hourly forecasting of pv power output, *IEEE Trans. Sustain. Energy* 5 (3) (2014) 917–926.
- [20] S. Manandhar, S. Dev, Y.H. Lee, S. Winkler, Y.S. Meng, Systematic study of weather variables for rainfall detection, in: *Proc. International Geoscience and Remote Sensing Symposium, IGARSS*, 2018.
- [21] P. Street, Dataport, Pecan Street Inc, 2018. <https://dataport.cloud/data/interactive>.
- [22] C.A. Smith, The Pecan Street Project: Developing the Electric Utility System of the Future, Ph.D. thesis, Citeseer, 2009.
- [23] B. McCracken, K. Rábago, M. Webber, Pecan Street Project Smart Grids and Austin's Energy Future, University of Texas at Austin: Environmental Science Institute, Austin, TX.
- [24] M. Kottke, J. Grieser, C. Beck, B. Rudolf, F. Rubel, World map of the köppen-geiger climate classification updated, *Meteorol. Z.* 15 (3) (2006) 259–263.
- [25] KNMI, Koninklijk Nederlands Meteorologisch Instituut, KNMI. <https://projects.knmi.nl/klimatologie/uurgegevens/selectie.cgi>, 2019.
- [26] R.A. Fisher, *Statistical Methods for Research Workers*, Genesis Publishing Pvt Ltd, 2006.
- [27] S. Raschka, V. Mirjalili, *Python Machine Learning*, Packt Publishing Ltd, 2017.
- [28] S. Diamond, S. Boyd, Cvxpy: a python-embedded modeling language for convex optimization, *J. Mach. Learn. Res.* 17 (1) (2016) 2909–2913.
- [29] A.J. Smola, B. Schölkopf, A tutorial on support vector regression, *Stat. Comput.* 14 (3) (2004) 199–222.
- [30] M. Rana, I. Koprinska, V.G. Agelidis, 2D-interval forecasts for solar power production, *Sol. Energy* 122 (2015) 191–203.
- [31] J.G.S. Fonseca, T. Oozeki, T. Takashima, G. Koshimizu, Y. Uchida, K. Ogimoto, Use of support vector regression and numerically predicted cloudiness to forecast power output of a photovoltaic power plant in Kitakyushu, Japan, *Prog. Photovoltaics Res. Appl.* 20 (7) (2012) 874–882.
- [32] L. Breiman, Random forests, *Mach. Learn.* 45 (1) (2001) 5–32.