HAL
open science

Bayesian Parameterisation of a Regional Photovoltaic Model - Application to Forecasting

Yves-Marie Saint-Drenan, Stephan Vogt, Sven Killinger, Jamie M Bright, Rafael Fritz, Roland Potthast

To cite this version:

Yves-Marie Saint-Drenan, Stephan Vogt, Sven Killinger, Jamie M Bright, Rafael Fritz, et al.. Bayesian Parameterisation of a Regional Photovoltaic Model - Application to Forecasting. Solar Energy, 2019, 188, pp.760-774. 10.1016/j.solener.2019.06.053 . hal-02174688

HAL Id: hal-02174688

https://hal.science/hal-02174688

Submitted on 5 Jul 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Bayesian Parameterisation of a Regional Photovoltaic Model - Application to Forecasting

Yves-Marie Saint-Drenan[a,*], Stephan Vogt[b], Sven Killinger[c], Jamie M. Bright[d], Rafael Fritz[b], Roland Potthast[e]

[a] *MINES ParisTech, PSL Research University, O.I.E. Centre Observation, Impacts, Energy, 06904 Sophia Antipolis, France*
[b] *Fraunhofer Institute for Energy Economics and Energy Systen Technology (IEE), 34119 Kassel, Germany*
[c] *Fraunhofer Institute for Solar Energy Systems (ISE), 79110 Freiburg, Germany*
[d] *Fenner School of Environment and Society, The Australian National University, 2601 Canberra, Australia*
[e] *Deutscher Wetterdienst (DWD), Offenbach, Germany*

Abstract

Estimating and forecasting photovoltaic (PV) power generation in regions——e.g. the area controlled by the transmission system operator (TSO)——is a requirement for the operation of the electricity supply system. An accurate calculation of this quantity requires detailed information of the installed PV systems within the considered region; however, this information is not publicly available making forecasting difficult. Therefore, approximating the undefined PV systems information for use in a PV power model (parameterization) is of critical interest. In this paper, we propose a methodological approach for parameterization using time series of aggregated PV power generation. A Bayesian approach is used to overcome the significant number of unknown parameters in the problem. It regularizes the linear system by imposing constraints on deviations from an initial-guess and covariance matrices; the initial guess uses available statistical distributions of PV system metadata. The performance of the proposed forecasting approach is evaluated using estimates of the regional PV power generation from three TSOs and meteorological data from the IFS forecast model (ECMWF). The proposed forecasting approach without the Bayesian parameterization has RMSE of 3.90%, 4.25% and 4.64%, respectively; including the Bayesian approach gives RMSE of 3.82%, 4.23% and 4.51%. For comparison, we also deployed a multiple linear regression which gave RMSE of 3.89%, 4.12% and 4.54%; however, there are considerable downsides to such an approach. Our approach is competitive with TSO forecasting systems despite using far fewer input data and simpler implementation of NWP prediction. This is particularly promising as there are many avenues for future development.

Keywords: PV system characteristics, regional PV power, forecast, Grid integration, inverse problem

*Corresponding author
Email address: `yves-marie.saint-drenan@mines-paristech.fr` (Yves-Marie Saint-Drenan)

1. Introduction

1.1. Background and motivation

With more than 400 GWp of installed photovoltaic (PV) capacity globally (IEA, 2018), the integration of the large amounts of solar energy in the electricity supply system is fundamental for modernization and maintaining grid reliability. The accurate estimation of power generated by a fleet of decentralized PV systems (hereafter referred to as regional PV power generation) is crucial at several stages of energy supply and network operations.

The objective of regional PV power estimates is to replicate the actual behaviour of the aggregated power production from all unknown PV systems installed in a given area; this can take advantage of all available information (power production measurements and/or PV system meta-data). Such systems have been described by Lorenz et al. (2011) or Schierenbeck and Graeber (2010). Estimation is made difficult because only a minority of systems continuously report their generation and few PV systems make their measurements publicly available—a serious issue that is the core subject of numerous studies (Bright et al., 2017; Lorenz et al., 2011; Shaker et al., 2015; Schierenbeck and Graeber, 2010).

A prominent application requiring regional PV power generation estimates is in the online and ex-post PV power analysis for grid monitoring and balancing-group settlement (Amprion, 2019). Grid operators are responsible for the estimation of aggregated PV power produced in their control area, as well as for publicly releasing the estimates as is often mandatory by law. An example of time series of the PV power generation estimated by most European transmission system operators (TSOs) are found on the ENTSO-E website[1].

Another important application group is providing the day-ahead or short-term forecasts of regional PV power generation. Forecasts are essential for energy trading (or scheduling thermal power plants), planning needs for reserve power or mitigating possible network congestion, etc. Improving the accuracy of regional PV power forecast is key because it has a positive impact on the integration costs of RES as well as on the security of supply (Killinger, 2017). Since the actual value of the regional PV power generation remains unknown, forecasting error is typically evaluated against the aforementioned regional estimates as a reference. Hence, the true goal of regional PV forecasts is to accurately predict the estimates made by the grid operators. It would be logical if forecasting methodologies used identical information as is used for the estimates; unfortunately, the data and processes involved in the estimation of the regional PV power generation are typically confidential so forecast providers must evaluate the regional forecast without it.

Within these two critical applications, two sources of uncertainty must be addressed in order to improve regional PV power generation estimates that are applied to forecasts: (i) the uncertainty resulting from the weather prediction error, and (ii) the uncertainty due to a lack of information on the installed PV systems.

[1] https://transparency.entsoe.eu

Each source of uncertainty represents a considerable field of research. The goal of this paper is to address the second source of uncertainty by proposing a method to infer the parameters of a regional PV model from times series of the aggregated PV power generation. Achieving this goal would enable the forecasts access to otherwise absent data, which, as will be demonstrated, can significantly improve the estimation.

1.2. Related work

Regional PV power forecasting research is maturing; it has particularly gained increasing interest in recent years. The approaches in literature are distinguishable principally by the strategies used to overcome uncertainty arising from a lack of information about the installed PV systems. In this literature review, research on regional estimates and forecast have intentionally been considered in tandem as the same algorithms are conventionally used in both; for absolute clarity, our methodology produces an estimate of PV power generation, we then assess improvements when the new estimate is used in a solar forecast.

A first approach is to assume that the PV power measurements of all systems installed in a region are known a-priori. Thus, the regional PV power forecast can simply be obtained by summing the forecasts from each PV system. This method is detailed by Da Silva Fonseca et al. (2014) where it is evaluated together with other methods in a benchmark analysis. Though this approach can be very insightful, it is difficult to make operational for two key reasons: poor access to PV power measurements, and linear computational scaling with increasing number of installations.

Lorenz et al. (2008, 2011) and Schierenbeck and Graeber (2010) proposed a pragmatic solution to the two aforementioned issues. The aggregated regional PV power generation is estimated from only a subset of the PV installations, limited to the most representative systems. The regional estimate is then reconstructed from the subset by means of an upscaling method. In Da Silva Fonseca Junior et al. (2014) and Shaker et al. (2015), the optimal subset of reference PV systems are determined mathematically using data-reduction techniques. A prerequisite of this method is access to an archive of all PV power measurements—a condition rarely satisfied. In Lorenz et al. (2008) and Lorenz et al. (2011), the choice of the most representative reference PV systems is based on a-priori knowledge on the fleet of PV system installed in the region as well as on spatial considerations. Whilst this latter technique is better suited for an operational implementation, it requires access to PV power measurements from a large number of installations, as well as a good knowledge of the metadata of installed PV systems in the considered region.

None of the approaches described previously can be implemented when too few PV power measurement data is available. Another kind of model can be used when lack of data is the barrier. As described by Saint-Drenan et al. (2017), the principle of this alternative method is to simulate the PV power generation of a limited number of commonly occurring PV system metadata configurations (in regard to capacity, tilt and azimuth) using meteorological data. The regional PV power estimate is then obtained by a weighted sum of the simulated power values, the weights corresponding to the frequency of occurrence of the considered

configurations. The unknown weights can whether be evaluated on the basis of authors' experience Schubert (2012); Fonseca Junior et al. (2015) or on the basis of a statistical analysis of PV system metadata Saint-Drenan et al. (2017); Killinger et al. (2018). A drawback of this approach is that possible differences between the linear coefficients chosen for the regional forecast and those corresponding to the regional estimates may penalize the forecast accuracy. This error can be minimized using model output statistics (MOS) techniques, which correct model outputs based on the information gathered from previous forecasts (Wilks, 2011). It is preferable, however, to directly use coefficients avoiding systematic errors; this is analyzed later.

Systematic differences between regional forecast and estimates can be avoided through use of supervised statistical methods, whereby the parameters of the model are trained using estimates of the regional PV power generation. A first example of this kind of approach can be found in the benchmark analysis by Da Silva Fonseca et al. (2014), where a support vector regression is realized using weather data as the input and a time series of the aggregated PV power generation is the output. In their work, the high-dimensionality of the input data penalizes the efficiency of the approach. Da Silva Fonseca et al. (2014) proved this by observing a noticeable improvement by using principal component analysis (PCA) of the entire weather information and accounting for 90% of the explained variance. A drawback of these types of method is that it requires important amounts of training data to learn the dependency between the input weather information and time series of aggregated power. Furthermore, whilst certain weather variables may account for significant variance in the aggregated power, that same variable may not have the same impact in different climates; hence, training data would always be required. The efficiency of the learning phase can be improved by using hybrid models, whereby known physical dependencies are considered and unknown model parameters trained from historical data. This possibility is shortly described in the "aggregated power curve" method proposed by Nuno Martinez et al. (2018) for the generation of stochastic solar area power forecast scenarios. Unfortunately, insufficient implementation and model performance information is provided by the authors. To the best of our knowledge, no other research investigating the potential of hybrid regional PV model for forecasting applications exists. This presents a significant opening for further research.

1.3. Contribution

Considering the lessons and outcomes identified in the literature review, we present a clear need for further investigation on the use of hybrid models for regional PV power forecasting applications. The objective of this paper is, therefore, to propose a physical regional PV power model whose parameters can be robustly inferred from estimates of the aggregate PV power production and to provide first results on its performance for forecasting the regional PV power generation.

The major benefit of achieving this goal is that regional PV power can be theoretically forecast without the uncertainties from a lack of reported metadata on the installed PV systems. Thus, regional PV power forecasts would be more accurate.

section 2 is dedicated to the description of the regional PV power model that has been used for this work. We have chosen the same formulation as previously used in Schubert (2012), Saint-Drenan et al. (2017) and Nuno Martinez et al. (2018). This modelling approach is linear so that unknown parameters can be obtained using regression techniques as in Nuno Martinez et al. (2018). We have paid a particular attention to the implementation details of this method (choice of the reference configuration, spatial constraints) in order to limit the number of unknowns without impacting the modelling accuracy.

section 3 focuses on the estimation of the model parameters. While the unknown parameters can easily be found by a simple regression, preliminary experiments have shown that a regression yields very high or negative parameters, which are physically meaningless and very sensitive to small variations in the training data set. These first observations, that contradict physical expectations, result from the "ill-conditioned nature" of the problem, which will penalize the power estimation accuracy and ultimately the forecast accuracy in application. To address this, we instead infer the parameters with a Bayesian method, which is a standard approach in inverse modelling. An additional benefit from a Bayesian approach is the integrating of an initial parametrization state such that previously defined iterations can be exploited to improve robustness.

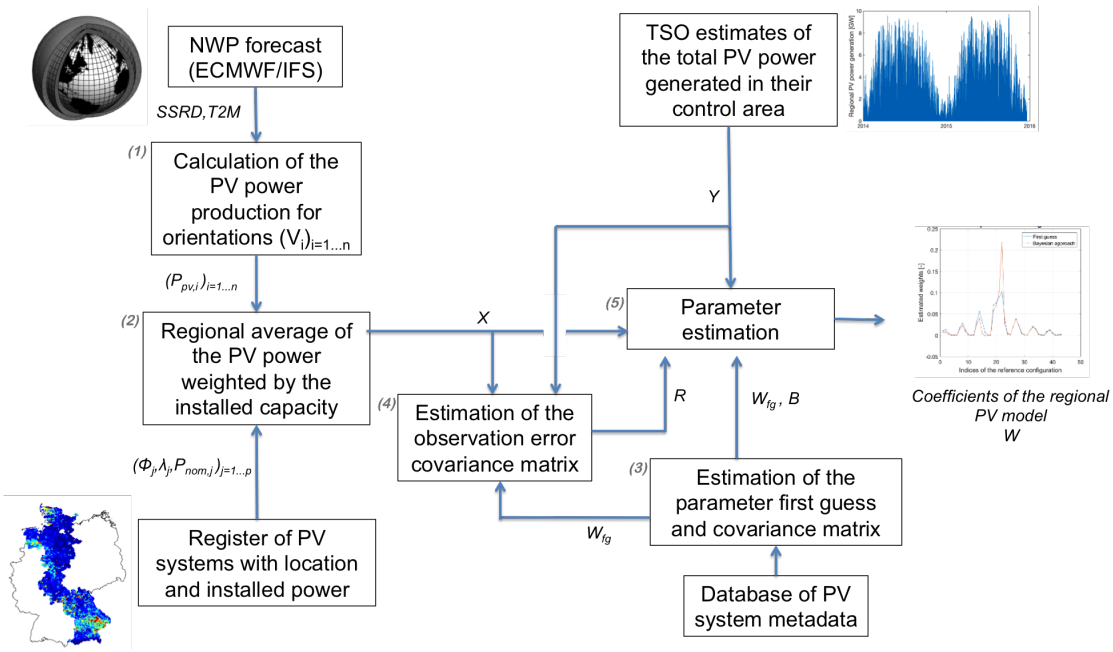
section 4 contains the results obtained from the proposed methodology in application to forecasting. One year of regional PV power generation of three German TSOs is improved using Bayesian parameterisation before being applied to day-ahead forecasting using corresponding weather forecasts taken from the IFS numerical weather forecast model. The benefit resulting from a Bayesian approach are quantified and the performances of the obtained forecast are compared to alternative forecasting approaches.

Finally, a discussion on the potential of the proposed method and concluding remarks are given in section 5.

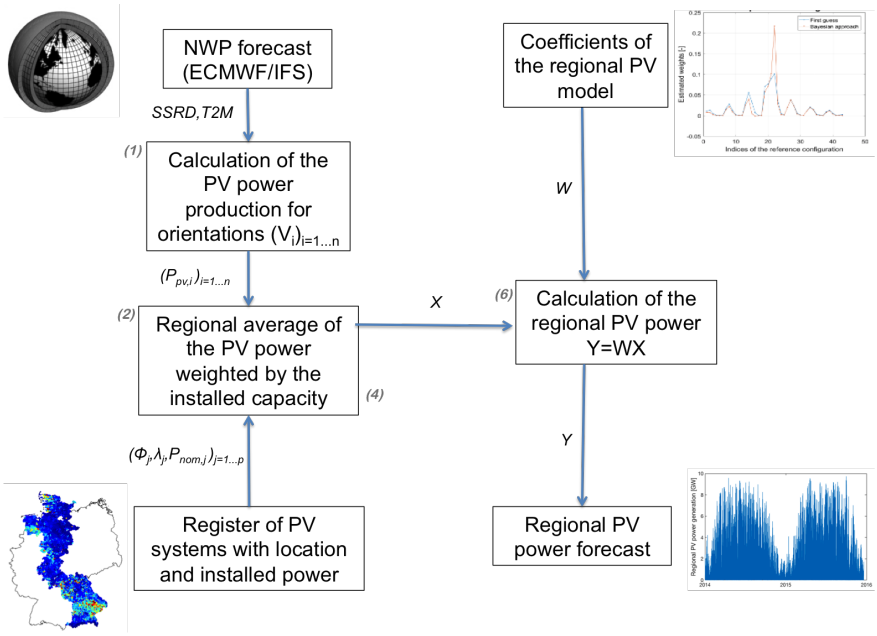
A flow chart summarizing our approach is given in Figure 1. This figure is referred to throughout this work to guide the reader on the different calculation steps of our method. Figure 1 is composed of two parts. The first part illustrates how the parameters of the regional model are inferred from the time series of the aggregated PV power production. The second part illustrates how the estimated parameters can be used to calculate regional PV power production using NWP data.

2. Regional PV power model

In this section, the regional PV power model used in this work is described in section 2.1. It is a linear model that calculates regional PV power generation from meteorological data. A set of reference PV system configurations is needed; the selection process is described in section 2.2. The loss of model accuracy resulting from our chosen set of configurations is then analyzed in section 2.3. In section 2.4, we describe the spatial constraints used to limit the number of unknowns due to the size of the regions. Finally, the



(a) Parameter estimation



(b) Calculation of the regional PV power forecast using the estimated parameters

Figure 1: Flowcharts summarizing (a) the calculation steps for the estimation of the parameters of the regional model, and, (b) calculation of the regional Pv power from NWP data using the estimated parameters.

modelling approach is summarized and matrix notation is introduced in section 2.5.

2.1. Description of the regional model

The regional PV model used in this work is the same as in Saint-Drenan et al. (2017); it is very similar to the approaches of Schubert (2012), Fonseca Junior et al. (2015) and Nuno Martinez et al. (2018). The model is based on the idea that, instead of simulating each PV system individually, power production from only those PV systems representing common PV system metadata configurations within a given region are simulated. They are then scaled up to the total capacity within the region. This approach represents a significant improvement in computational efficiency.

The regional power generated by a fleet of PV systems installed at location x can be expressed as the sum of the power values calculated with characteristics V_i multiplied by the share w_i of the total capacity of systems having the characteristics V_i. The simulated power is normalized by the corresponding peak power and the weighted sum scaled to the actual value of the installed capacity at location x. The regional PV power generation can then be expressed as:

$$P_{PV,t,region} = \underbrace{\sum_{x \in region} \underbrace{P_{installed,x}}_{\text{installed PV power}} \times \underbrace{\sum_{i=1}^{n_{ref}} (\underbrace{w_{i,x}}_{\text{weight}} \times \underbrace{f_{PV}(x, t, G_{x,t}, T_{2m,x,t}, V_i)}_{\text{power gen. of a ref. PV system}})}_{\text{aggregated PV power at location x } (P_{PV,t,x})}}_{\text{aggregated PV power for the region}}, \quad (1)$$

where:

- t is the time
- x represents the different locations within the considered region
- $P_{PV,t,region}$ is an estimate of the aggregated power produced by all PV systems in the considered region at time t $[W]$
- $P_{installed,x}$ is the installed PV power at location x $[W_p]$
- $P_{PV,t,x}$ is an estimate of the aggregated power produced by all PV systems at location x at time t $[W/W_p]$
- $w_{i,x}$ is the weight of the ith reference configuration at location x [-]
- $f_{pv}(...)$ is a function representing the single PV system model used to calculate the normalized PV power $[W/W_p]$
- V_i is a vector with the configuration parameters of the i_{th} reference PV system
- $G_{x,t}$ is the global horizontal irradiation at location x and time t $[W/m^2]$
- $T_{2m,x,t}$ is the air temperature at x and t $[°C]$

The function f_{PV} in Eq. 1 represents a single PV system model that needs to be chosen beforehand. This function corresponds to the step (1) in the flow chart displayed in Figure 1. Saint-Drenan et al. (2017) demonstrated that a simple model with a limited number of input parameters performs well in regional

applications; thus, we select the same model for the present work. The calculation steps of this model are illustrated in Figure 2 and a detailed description can be found in (Saint-Drenan, 2015).

With the chosen model, the set of characteristics V_i is only composed of the azimuth angle $\alpha_{M,i}$ and module tilt angle $\gamma_{M,i}$. Another important model parameter that is not explicitly considered here is the total efficiency of the PV system. Though large variations of the efficiency may be observed among PV systems (Killinger et al., 2018), we decided to use a constant. That said, variations of the system total efficiency are implicitly considered in our regional model through the weights w_i. Therefore, these weights not only reflect the distribution of capacity across all orientations, but also account for the efficiency of the PV systems.

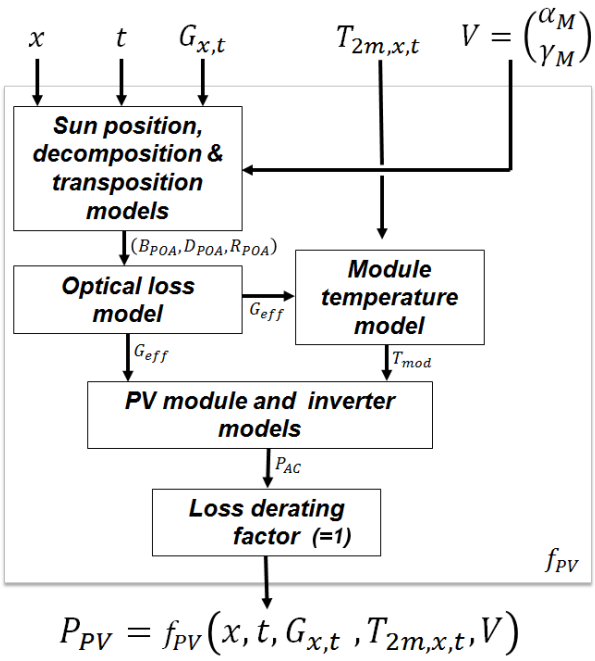


Figure 2: Flowchart of the single PV system power model.

2.2. Choice of the reference orientation combinations

In the single PV system model (see section 2.1), orientation of a PV system is defined by two parameters: azimuth and tilt angles. It is obvious that the power production simulated for two close module orientations are highly correlated, and so we do not necessarily need every combination of tilt and azimuth, only a representative set. Therefore, a smaller subset of orientation combinations that allows an accurate calculation of the aggregated PV power generation is needed. We call this subset the 'reference orientation combinations'. Selecting too many combinations could lead to numerical issues during the search of the model parameters due to the high number of unknowns and their co-linearity. Therefore, we use a regular

grid of tilt and azimuth angles with a coarse resolution of 15° in both dimensions to define the reference orientation combinations.

To further limit the number of unknown parameter of our model, we exclude uncommon combinations. Based on Saint-Drenan (2015) and Killinger et al. (2018), we limit the azimuth angles values between -45 and 45°, and tilt angles between 0 and 45°. Ultimately, we represent all systems by 22 reference orientation combinations.

2.3. Uncertainty arising from the reference orientation configurations

In the previous section, we opted for a coarse grid of reference orientation combinations. Though motivated by numerical considerations, the question arises of 'how much loss of accuracy results from from this choice?' To answer this question, we calculate the error when Eq. 1 is used to calculate the power production of a single PV system with arbitrary orientation. As errors within the regional model are expected to balance out with a increasing number of PV systems, the consideration of a single PV system should provide a worst-case indication of the error. To demonstrate, we attempted to rebuild the power production of an arbitrary orientation using our set of 22 reference orientation combinations. For this purpose, the power production was calculated using single PV power model for the 22 reference orientation combinations as well as for the single arbitrary test orientation. We used one year of 15-min global horizontal irradiation measurements and air temperature measurements from Fraunhofer IWES in Kassel. The weights of the linear model were then calculated by a multiple linear regression between the time series of the calculated power for the test orientation and the time series from the reference orientations.

The underlying assumption of Eq. (1) that the output of a PV system with an arbitrary orientation can be assessed by a linear combination of the outputs of different PV systems corresponding to reference orientations is illustrated in the left panel of Figure 3: the squares represent each reference orientation; the circle represents the arbitrary test orientation; the numbers in boxes and the lines show the regression coefficients. In the right panel of Figure 3, the power values calculated with the physical model are displayed against the linear combination of reference power values. The difference between the two power values are negligible.

This procedure is repeated for all integer value combinations of the azimuth angle (between -45 and 45°) and the tilt angle (between 0 and 45°). The resulting root mean square difference (RMSD) of the residual of the regression are represented by colored squares in Figure 4.

All RMSD values are less than 2.10^{-4} W/W_p; this analysis shows that the power values calculated with the 22 reference orientation combinations is coarse enough to reconstruct the power for an arbitrary orientation (as long as this orientation lies in the domain covered by the reference orientations). We observe peaks in RMSD (though still small) at around 40° tilt and between 45° to 30° east and west. Since such orientations are infrequent in Germany, we consider these greater RMSD values to be of insignificant impact.

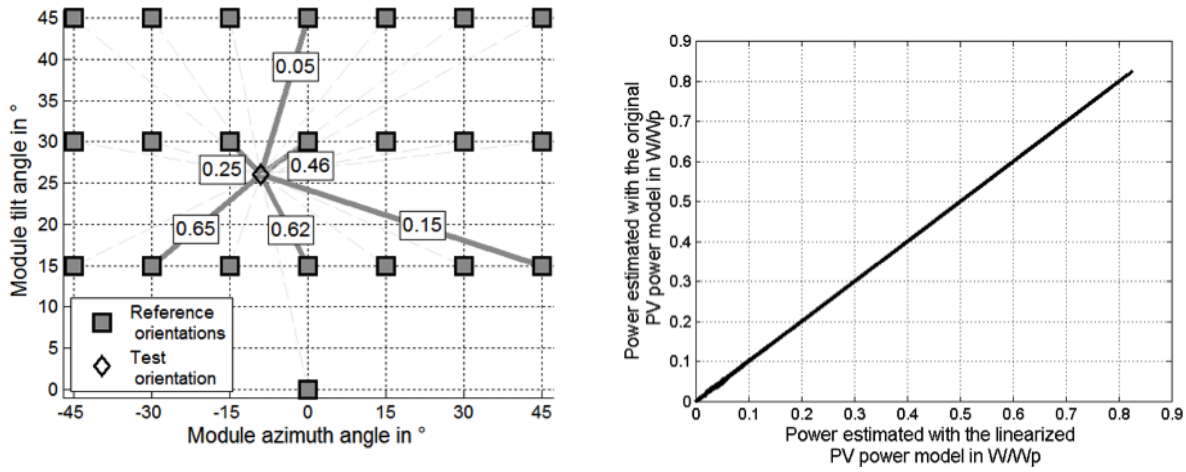


Figure 3: Left panel) Illustration of the approach used to estimate the PV power at a given orientation as derived from the PV power value estimated with the reference orientations using a linear approach. The reference and the test orientations are represented by grey squares and a circle, respectively. Weighted coefficients are indicated in each box. Small coefficients (≤ 0.01) are not displayed. Right panel) Scatter plot of power values evaluated with the physical model against power values evaluated with the linear approach.

Control area or region	**Number of PV systems**	**Size of the region**
TenneT	646 968	140 500 km^2
50Hertz	125 696	109 000 km^2
Amprion	437 622	73 100 km^2
TransNetBW	281 420	34 600 km^2
Germany	1 491 706	357 168 km^2

Table 1: Number of installed PV systems in the four German control areas in August 2014.

2.4. Spatial constraints

At this stage, there is a total of $n_{ref} \times n_{loc}$ unknowns in our regional model, where n_{ref} is the number of reference orientation combinations (22 orientations selected in the previous section) and n_{loc} is the number of locations considered in the regions. Considering the size of a control area and the great number of PV systems installed in a control area (Table 1), it is necessary to add spatial constraints to the problem in order to limit the number of unknowns.

Given that the unknowns are inferred from regionally averaged PV power generation, it is unrealistic to expect that local information can be retrieved from the method. In addition, the possible impact from errors at the local scale are balanced when upscaled to a regional level; thus, they are deemed insignificant.

Therefore, we are mainly interested in large scale spatial trends that may have a significant impact on the aggregated regional PV power generation estimate. It is possible to add spatial constraints by considering

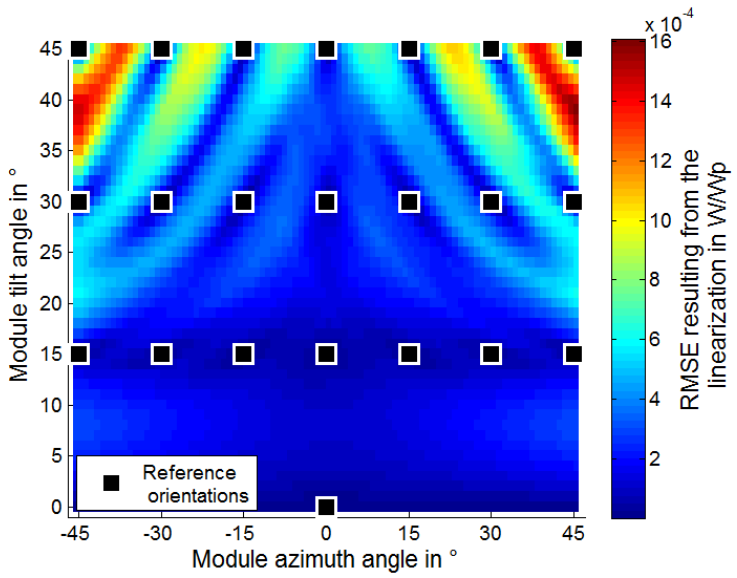


Figure 4: Root mean square difference (RMSD) between the PV model output and the corresponding value obtained with the 22 reference orientation combinations using a multilinear regression.

large scale regional trends; approaches exist such as spatial regularization techniques or the use of spatial parametric model. We selected the simplest and most pragmatic method—dividing a region into a reasonable number of sub-regions and assuming that unknown parameters are constant in each sub-region.

Regions are divided into sub-regions using a k-mean clustering algorithm on the coordinates of PV systems installed in the greater region. Two examples are illustrated in Figure 5, where the control area of TenneT is divided into 3 and 5 sub-regions, respectively. In this figure, light grey dots represent the known installed PV systems and dark squares are centroids of the clusters (sub-regions). The borders between different sub-regions are displayed by blue lines.

If n_{SbRg} is the number of sub-regions, the number of unknown of the linear system is now equal to $n_{ref} \times n_{SbRg}$, for example $n_{ref} = 22$ and $n_{SbRg} = 2...5$ tractable.

2.5. Summary

The weights $w_{i,x}$ are assumed constant within each sub-region R_j. Thus, the model formulation in Eq. 1 can be simplified by introducing the smaller set of w_{i,R_j} for each configuration i and region R_j. The weights $w_{i,x}$ are related to the weights w_{i,R_j} by the relationship $w_{i,x} = w_{i,R_j} \forall x \in R_j$. With this new variable, Eq. 1 is written:

$$\underbrace{P_{PV,t,region}}_{\text{Y[t,1]}} = \underbrace{\sum_{R_j} \sum_{i}}_{\sum_k} \underbrace{\left(w_{i,R_j}\right)}_{\text{W[k,1]}} \times \underbrace{\left(\sum_{x \in R_j} P_{installed,x} \times f_{PV}(x, t, G_{x,t}, T_{2m,x,t}, V_i). \right)}_{\text{H[t,k]}} \tag{2}$$

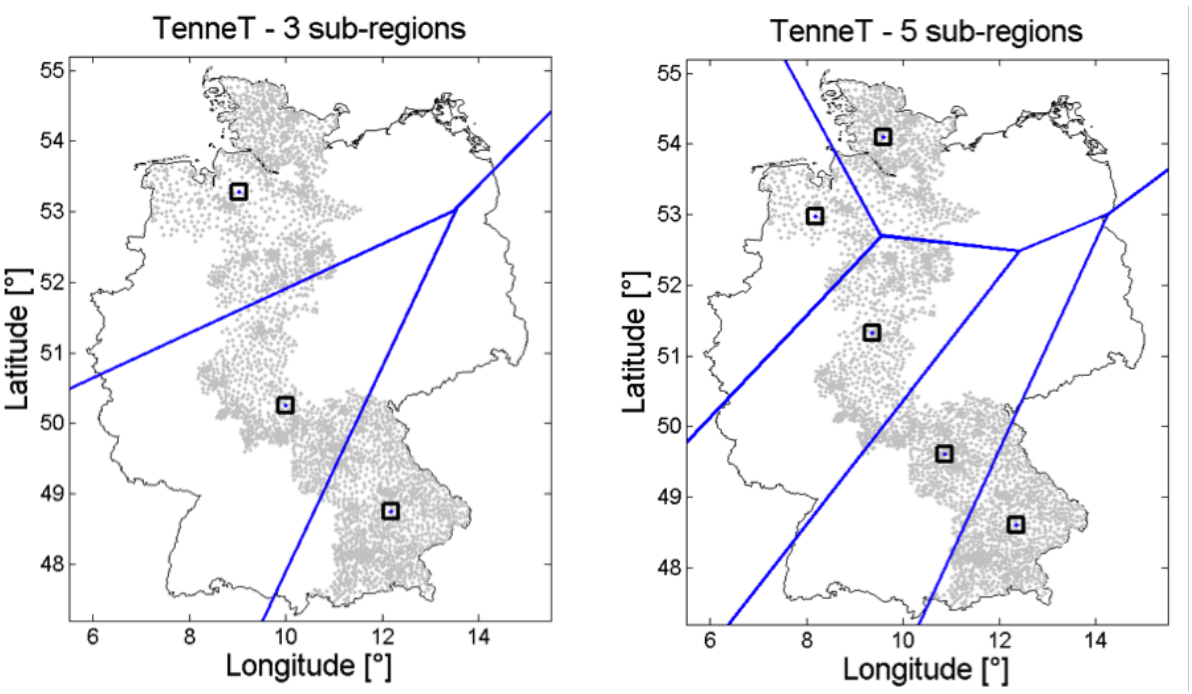


Figure 5: Illustration of the split of the TenneT control region into 3 sub-regions (left map) and 5 sub-regions (right map) using k-mean clustering. The centroids of each sub-regions are marked by a black square and the limits are indicated with blue lines.

For the implementation of Eq. 2, PV system installed capacity $P_{installed,x}$ and meteorological information $(G_{x,t}, T_{2m,x,t})$ are needed for each location x of the considered region. With this information, the right-side of Eq. 2 ($H[t,k]$) can be calculated for each sub-region R_j. This quantity corresponds to the weighted sum of the simulated power values for a configuration V_i, the weights being the installed capacities. The unknown of the problem are the weights w_{i,R_j} for each reference configuration V_i and sub-region R_j.

It is convenient to express Eq. 2 in matrix notation $\boldsymbol{Y} = \boldsymbol{HW}$. The column vector $\boldsymbol{Y}$ contains aggregated PV power generation in a given region at different time steps t. The vector $\boldsymbol{W}$ contains the unknowns of the problem w_{i,R_j}, and H is a matrix containing the sum of the simulated PV power values for each combination of sub-region and reference orientations (column) and time steps (row). For the sake of clarity, the relationship between the summation and matrix notations are indicated by underbraces in Eq. 2. In this illustration, k is an index that screens all combinations of reference orientations i and sub-regions j.

The above described summation of the simulated PV power weighted by the installed capacity to evaluate the matrix X corresponds to the step (2) of the flow chart given in Figure 1.

3. Estimation of the model parameters

3.1. Approach

The problem described in section 2.5 is inverse: we start with results (aggregate PV power generated within the region) then calculate the cause (weights for all PV systems with the reference orientation combinations). This is the opposite to previous methodologies (Nuno Martinez et al., 2018; Saint-Drenan et al., 2017) where the authors start with the cause (the distribution of the PV systems according to the region and reference orientation combinations) and then calculate the result. Inverse problems feature heavily in literature (Nakamura and Potthast, 2015; Colton and Kress, 1997; Engl et al., 2000). They are common in many research fields, especially in meteorology: the retrieval of cloud optical characteristic, the assimilation of measurements in numerical weather prediction models, to name a few.

Solving inverse problems is non-trivial as they are typically ill-posed. Among the three Hadamard conditions for a well-posed problem (Hadamard, 1902)—existence, uniqueness, and stability of the solution—the conditions of uniqueness and stability are often violated. Particularly with a great number of unknown parameters, an observation can be explained by several causes (violation of the uniqueness condition). In our application, multicollinearity of the input data poses an issue (see section 2.3). The linear dependency among regressors is reflected in the solution space, where many possible combinations of the solution vector components lie within a narrow valley (in high dimension) all positioned very close to the minimal residual sum of squares. In such cases, changes in the model output can instead be obtained by changes in the model input parameters.

Small perturbations in the input data can bring about noticeable changes in the solution (violation of the stability condition). Such problems can be addressed by using regularization techniques; for example, the generalized Tikhonov regularization, whereby the deviation of the solution from the initial guess is penalized (Tikhonov, 1963). This approach requires the choice of regularization parameters that balance the respective effects of the error and regularization terms (Nakamura and Potthast (2015), Chapter 3.1.6). There are many techniques for motivating the choice of the regularization parameter; however, selection remains a non-trivial and delicate issue. Alternatively, regularization parameter selection can be entirely avoided with a Bayesian approach.

In the Bayesian framework (see for example Nakamura and Potthast (2015) chapters 4.2 and 5.6, Crisan and Bain (2009) or Crisan et al. (2014)) our goal is to find the set of parameters $\boldsymbol{W}$ giving the highest posterior probability $P(\boldsymbol{W} \mid \boldsymbol{Y})$ given $\boldsymbol{W}$. To this end, we use the well known Bayes law:

$$\underbrace{P(W \mid Y)}_{\text{posterior}} \propto \underbrace{P(W)}_{\text{prior}} \underbrace{P(Y \mid W)}_{\text{likelihood}} \tag{3}$$

We assume that the prior $P(W)$ can be approximated by a Gaussian distribution with mean value $\boldsymbol{W_{fg}}$ and covariance matrix $\boldsymbol{B}$. Similarly for the likelihood $P(Y \mid W)$, we take a Gaussian distribution with

zero mean and covariance matrix $\boldsymbol{R}$. The covariance matrix $\boldsymbol{B}$ can be interpreted as a quantification of the possible variations of the resultant vector around the first guess; the second covariance matrix $\boldsymbol{R}$, a quantification of the uncertainty of the observations $\boldsymbol{Y}$.

With these notations, the posterior probability can be written:

$$P(W \mid Y) \propto exp\left(\frac{1}{2}(W - W_{fg})^T B^{-1}(W - W_{fg})\right) \times exp\left(\frac{1}{2}(HW - Y)^T R^{-1}(HW - Y)\right). \quad (4)$$

Given than maximizing the likelihood is equivalent to minimizing the logarithm of the above expression (Freitag and Potthast, 2013), the desired solution W_{opt} is to minimize the following functional:

$$J(W) = \frac{1}{2}(W - W_{fg})^T B^{-1}(W - W_{fg}) + \frac{1}{2}(HW - Y)^T R^{-1}(HW - Y). \quad (5)$$

The solution that minimizes the above cost function J has zero gradient ($\nabla_W J = 0$). This condition allows the explicit determination of the solution to Eq. 5:

$$W_{opt} = W_{fg} + \left[B^{-1} + H^T R^{-1} H\right]^{-1} \times \left(H^T R^{-1}\right)(Y - HW_{fg}) \quad (6)$$

This relationship, which corresponds to the expression of the generalized Tikhonov regularization, can now be used for estimating the weights that correspond to the different reference orientation combinations. To this end, the estimation of the initial guess and covariance matrices ($\boldsymbol{R}$ and $\boldsymbol{B}$) is achievable—detailed in the following subsections. It is illustrated by the step (5) in Figure 1.

3.2. Determination of the initial guess

As mentioned previously mentioned, the unknown parameters $\boldsymbol{W}$ can be interpreted as the distribution of PV systems according to the different possible orientations and to each sub-region R_j. This interpretation can be exploited to evaluate the initial guess W_{fg}.

The same approach to evaluate our initial guess was presented by Saint-Drenan et al. (2017). A database including module orientation angles for more than 20,000 systems is used to evaluate the share of the total capacity corresponding to each of the reference orientation combinations. When evaluating the initial guess, we neglected the potential geographical dependence of the parameters of our regional model. The distributions were thus evaluated using all PV systems in the database regardless of their location in Germany. If sub-regions are later considered, we assume that the same distribution can be used as an initial guess for each sub-region. Regions can and do present distinct differences in orientation; when considering a larger aggregate statistic, these nuances can be ignored—Killinger et al. (2018) (fig. 3) visually demonstrates a distinct north-south division in PV system orientation for France. In Figure 6, the components of the first guess vector are represented by squares that are coloured as a function of the module azimuth and tilt angles. The statistical analysis aiming at the estimation of the first guess is represented by the step (3) in Figure 1.

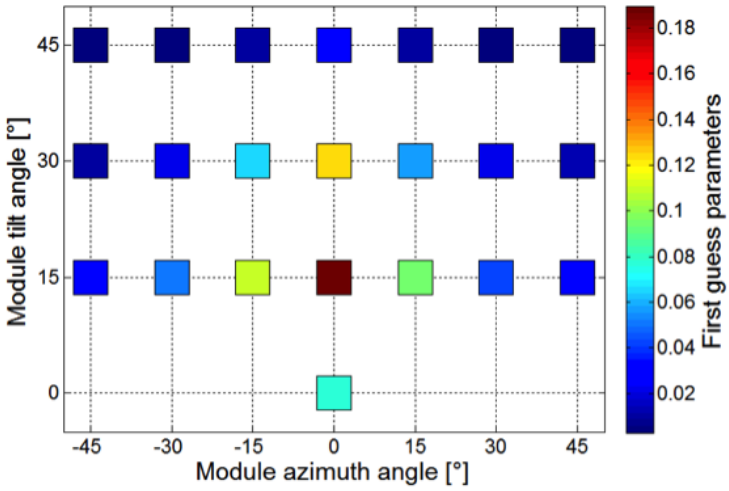


Figure 6: Values used for the initial guess (colour of the squares) as a function of the module azimuth and tilt angles.

3.3. Estimation of the background covariance matrix

The background covariance matrix $\boldsymbol{B}$ quantifies the expected dispersion of the parameter vector around the initial guess W_{fg}. In order to evaluate this distribution, we generate a set of perturbed first guess vectors using the approach described in section 3.2:

$$\left\{ \left[\tilde{w}_1^{(p)} ... \tilde{w}_j^{(p)} ... \tilde{w}_{n_{ref}}^{(p)} \right]^T .p = 1...10000 \right\} \tag{7}$$

The perturbation uses only 1000 randomly chosen PV systems instead of the total dataset. Indeed, a random sub-sampling of the set of PV systems brings about variations of the original distribution of PV systems according to the reference orientation combinations. However, we assume that with a sufficiently small number of samples and a sufficiently large number of iterations, the range of possible values taken by the resulting distribution is accounted for. This approach is re-iterated 10,000 times to populate a set of possible solutions that are subsequently used for the estimation of the background matrix $\boldsymbol{B}$.

The set of randomly evaluated parameter sets is then used to calculate each component $b_{i,j}$ of the matrix $\boldsymbol{B}$ using the following relationship:

$$b_{i,j} = Cov\left(\left[\tilde{w}_1^{(i)} ... \tilde{w}_k^{(i)} ... \tilde{w}_{n_{ref}}^{(i)} \right]^T, \left[\tilde{w}_1^{(j)} ... \tilde{w}_l^{(j)} ... \tilde{w}_{n_{ref}}^{(j)} \right]^T \right). \tag{8}$$

All PV systems are considered in the covariance estimation so that the resulting matrix reflects the true covariance of the model parameters for Germany (note that, when applying this methodology to a different country/region, a representative covariance matrix is required). We assume that the same covariance holds true for each sub-region and that the covariance of two parameters in different regions is null.

The estimation of the background covariance matrix using a dataset of PV system metadata corresponds to the step (3) in Figure 1.

3.4. Estimation of the observation error covariance matrix

The observation error covariance matrix quantifies the uncertainty of the observation vector $\boldsymbol{R}$. We assumes that this error is time-independent so that $\boldsymbol{R}$ can be expressed as the product of an identity matrix with a scalar. We further assume that the variance of the error obtained with the initial guess $\boldsymbol{W_{fg}}$ evaluated over a training time period can be used as an approximation of that of the error variance. As a result, $\boldsymbol{R}$ can be evaluated as follows:

$$\boldsymbol{R} = Var\left(\boldsymbol{H}\boldsymbol{W_{fg}} - \boldsymbol{Y}\right) \cdot \mathbb{I} \tag{9}$$

The estimation of the $\boldsymbol{R}$ is subjected to several assumptions that can have a non negligible impact on the results. By choosing this simple and rough approach to gain insight on the benefit of the Bayesian approach applied to our problem, future work is needed to improve the estimation of this matrix. It is represented by the step (4) in Figure 1.

4. Results

This section is separated into four parts. The data are firstly introduced in section 4.1. The training period and validation procedure is described in section 4.2. Model performance looking particularly impacts from regularization and changing number of sub-regions is presented in section 4.3. Comparisons of the newly proposed forecasting approaches are compared to persistence, initial guess, and the TSO forecast in section 4.4.

4.1. Data

The first type of information needed for implementing our method is an estimation of the aggregated PV power produced in the considered area. In this work, we use TSO estimates of the aggregated PV power generation for three German control areas: TenneT, Amprion and 50Hertz. For each control area, time series of the estimated aggregated PV power generation as well as the installed PV system metadata (location, peak power and time of installation) are available. The time series of the regional PV power estimates have a temporal resolution of 15 min and are available for the years 2014 and 2015. A short description of these data provided by the German TSOs can be found in Saint-Drenan et al. (2017).

The second input required by our model is meteorological data and more precisely the global horizontal irradiation and air temperature. The most natural choice is to select the most accurate gridded data, as for example reanalysis and/or satellite data. As reported by Ineichen (2014) or Boilley and Wald (2015), these data have their own sources of error that can add to forecast error of the NWP data used for the prediction. We therefore chose the pragmatic option of using directly the NWP forecast data for the parameter estimation in a way to avoid double penalty error described above. Forecasts from the Integrated Forecast

System (IFS) model run by the European Center for Medium-Range Weather Forecast (ECMWF) are therefore used for the calculation of the PV power generation for each region and for each reference orientation (matrix $\boldsymbol{H}$). Hourly gridded 2m ambient temperature (2T) and solar surface radiation downwards (SSRD) are available at $0.125 \times 0.125°$ spatial resolution; they are collected, prepared and linearly interpolated to a 15 min temporal resolution corresponding to the PV power. Only forecast data with a lead time between 24 and 48h for the run starting at 00:00 (GMT) are used to evaluate day-ahead forecast.

Finally, we used the operational forecast published by the three TSOs on their website to benchmark the output of our model. These forecast are calculated by the TSOs using several forecasts of the regional PV power generation from private forecast providers. These individual forecasts are optimally mixed and calibrated using the reference PV production value.

In this work, we decided to show the potential of our approach using real-world operational data: the parameters of the regional model are estimated using TSO estimates of the PV power production and the model output is compared to TSO day-ahead forecasts. If this approach allows demonstrating the practical relevancy of our work, it has also some drawbacks: uncertainties in the TSO data [2] used for training and validation can bring about errors, that are impossible to differ from the actual error of our method. An alternative to avoid these undesired effects would have consisted in generating synthetic data. Yet, we preferred evaluating the performances of our method in real conditions because the motivation of the present work is strongly linked to the targeted application. As a consequence, it is important to bear the real-world settings of the validation in mind to properly interpret the results presented in this chapter.

4.2. Training and validation setup

An adaptive approach has been chosen to train the model (see Figure 7). The model parameters are evaluated and tested for the year 2015 on a monthly basis using 12 months of training data preceding each test time-period. This restrains various issues that could lead to differences in characteristics between training and testing data, such as regular improvement of NWP models, change of the TSO estimates (e.g. extension of the set of reference systems with time), or, new installed PV systems over time.

The presence of snow on PV systems results in negligible PV power generation that would not be appropriately represented in our proposed system without additional complexity. Since this effect is not considered, days marked by the presence of snow in Germany must be excluded from the training and testing data (see Figure 7). To do this, days with snow have been identified using snow height data at 200 German SYNOP stations collected from the OGIMET website (`www.ogimet.com`). The network of SYNOP station is coarse; we decided to exclude each day when snow is reported at more than one meteorological station as

[2] The major sources of uncertainty are the uncertainty of the regional estimates and the information on the installed PV capacity.

Figure 7: Graphical representation of the training and testing setup (upper plot) and number of days available for training and testing (lower plot).

a conservative measure. Should large NWP forecast errors be present at the training phase of the model, the whole approach will be unjustly penalized. To avoid this, all days were excluded from the training dataset only if at least one value of the first guess forecast error exceed $0.2W/W_p$. This latter criterion is not applied to the testing data set.

Finally, training and testing routines were conducted at three control areas with a varying number of sub-regions (1 to 5).

For each forecast considered in this work, all standard error metrics were evaluated using only daytime values; these results are presented in Table 2. In the two next sections, the performances of the forecast methods will be discussed with a focus on the RMSE values. Indeed, this metric is widely used by TSOs and forecast providers to assess the forecast accuracy as it well reflects the expectations on the performances of the forecasting methods for grid integration mechanism.

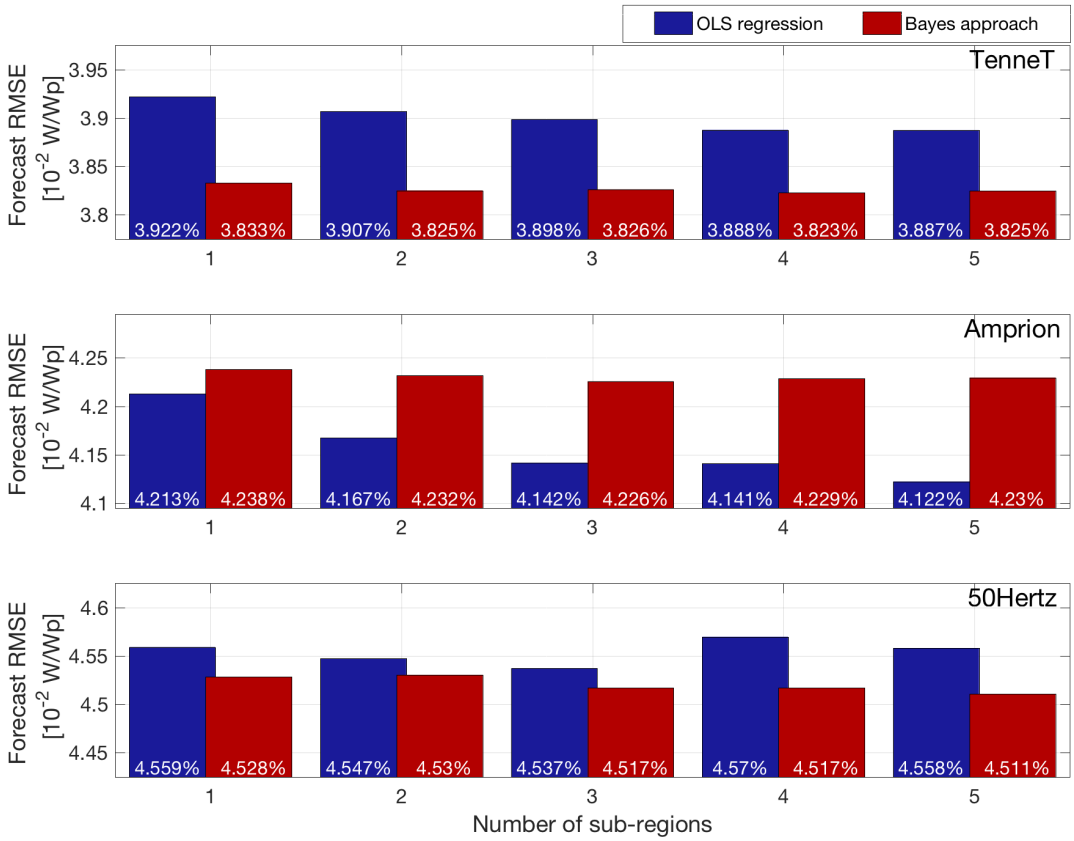


Figure 8: Effect of the number of sub-regions on the model performances for the three considered control areas (the limits of the y-axis have been scaled to illustrate best the differences between models).

4.3. Model performance impact from regularization and varying number of sub-regions

In Figure 8, RMSE values obtained with an ordinary least square regression (OLS) and the Bayesian method are displayed for 1 to 5 sub-regions. Considering Figure 8, it is remarkable that the two approaches yield very different results based on the data of the three transmission system operators. Disregarding the influence of regions and focusing only on the average differences between the OLS regression and the Bayesian method, we confirm that the added value of the regularization is not systematic. A beneficial and a moderate improvement of regularization can be observed for TenneT and 50 Hertz, respectively, as indicated by smaller RMSE values from the Bayes method; regularization brings about an increase of the RMSE values for Amprion. There are likely many causes for the observed differences; however, the most probable explanation is that the initial guess—and to a lower extent the covariance matrix B—computed for the whole Germany is not representative for the three control areas. The initial guess and the covariance matrix approach was motivated by the limited number of information about the PV systems installed in Germany. A calculation with a initial guess and covariance matrix $\boldsymbol{B}$ specific to each region would be

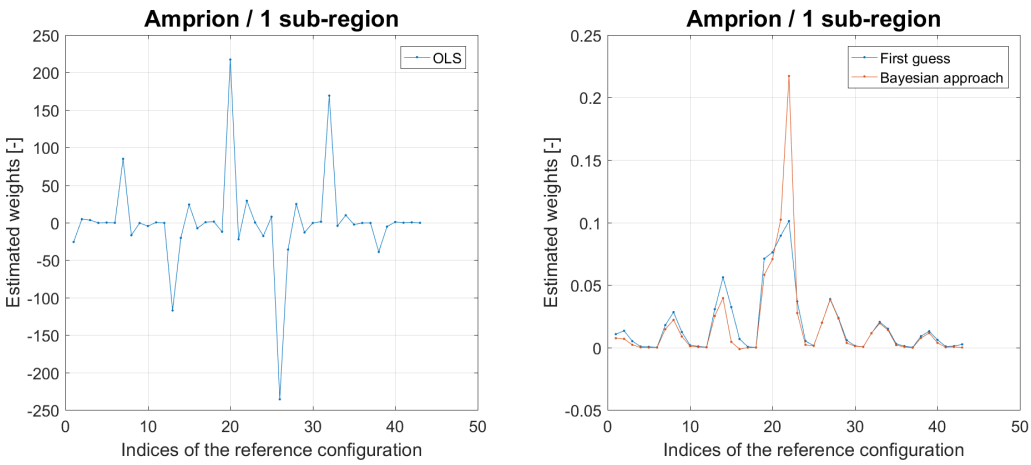


Figure 9: Comparison of the weights evaluated with the ordinary least-square regression (OLS) and those obtained with the Bayesian method with one sub-region for the Amprion control area.

of interest at a later data as soon as more information is available on the installed PV systems in the respective regions. Even though RMSE from the Bayes method is worse than that of the OLS regression in Amprion, the resulting parameters are still more robust due to the nature of a Bayes methodology. This is observed in Figure 9 where the parameters obtained from the OLS regression are compared to the regularized counterpart. While the Bayes method parameters closely resemble the actual values, those obtained through OLS regression can be very high and even negative—physically impossible. Under these circumstances, if testing data were to exist outside the parameterized domain from the training data, it must be accepted that a forecast trained with OLS regression may yield meaningless results due to the collinearity of the input features —this will not happen with the Bayesian method. Physically meaningful parameters are preferable –even at the cost of a lower performance– in operational context, where a reliable forecast system is needed.

The evolution of the RMSE with increasing number of sub-regions is also very different across the three main regions. While increasing the number of sub-regions results in a small but steady RMSE decrease with the Bayes method for the three control areas, the effect of the number of regions on the performances of the OLS method is very different from one control area to another. In the TenneT and Amprion control areas, it is observed that the RMSE decrease for OLS regression is relatively more consistent and pronounced. By contrast in the 50 Hertz control area, RMSE from OLS regression decrease from 1 to 3 sub-regions, but increase again thereafter; this is presumably the result of a lack of generalization from too many degrees of freedom with more sub-regions (overtraining). With too many parameters, the model has a tendency to learn information that corresponds too closely to the training data set (such as rare events or even noise); therefore, it fails to generalize the trained dependencies to testing data. If inferring model parameters for 1 to 3 sub-regions can appear less considering the numerous PV systems installed in the different regions, it should be pinpointed that the spatial differentiation is made using only one time series of the aggregated

PV power production. It can therefore be expected that modifying the approach to integrate more data (DSO aggregated power, time series of single PV systems, dataset of PV metadata...) would enable a better spatial characterization of the unknown fleet of PV systems.

4.4. Benchmarking model performance against alternative forecasts

In this section, the forecasting performances from the Bayesian method is compared to alternative forecasts methods in order to evaluate any added value. We evaluated the performance of 5 different forecasting approaches at each of the three control areas. RMSE results are displayed in Figure 10.

The first comparison is against smart persistence, whereby the current day's TSO estimate is used as the forecast for the following day. The second comparison is from the initial guess (without training phase), which corresponds to the approach described in Saint-Drenan et al. (2017). The performances of the Bayesian method and OLS regression with 1 to 5 sub-regions are also considered in this evaluation for additional insight. The final comparison is against the publicly available day-ahead forecasts from the German TSOs.

It is generally considered good practice to include persistence forecast in a benchmark as it shows the added value of more advanced method (Yang et al., 2017). Persistence can yield remarkable results with stable weather conditions over several days, however leads to significant errors under variability. For day-ahead prediction, it is well established that NWP-based forecasts outperform persistence forecasting—Figure 10 corroborates this.

The exceptional performance of the initial guess, Bayesian method and OLS regression forecasting at Amprion is not representative and is now a subject of discussion with the German TSO; clearly the published forecast not the best forecasting approach, even though it has been made public by Amprion. Similar observations were already made by Saint-Drenan et al. (2017) in a previous work. We believe it would be a poor representation of the described approach to consider Amprion in our analysis as there are evidently issues with the published Amprion data; as such, Amprion data are not considered in the present discussion.

Comparing the Bayesian method performance against the initial guess quantifies the benefits of training the parameters on historical data. We clearly demonstrate that including the training stage reduces the RMSE by 0.08 and 0.12 10^{-2} W/W_p for TenneT and 50 Hertz, respectively, representing relative improvements of 2.05 and 2.59%—a small but clear improvement. Ultimately, training the model parameters offers real added value to the forecasting skill. We also note that the initial guess is already a reasonable approximation of the true model parameters.

Comparing to the RMSE from the initial guess, the Bayesian method and OLS regression corroborates our initial assumption that the parameters of the regional model can be inferred from the aggregated PV power generation. For TenneT, OLS regression RMSE is greater the Bayesian method and initial guess RMSE results. This indicates that Bayesian regularization avoids issues from overtraining. For 50 Hertz,

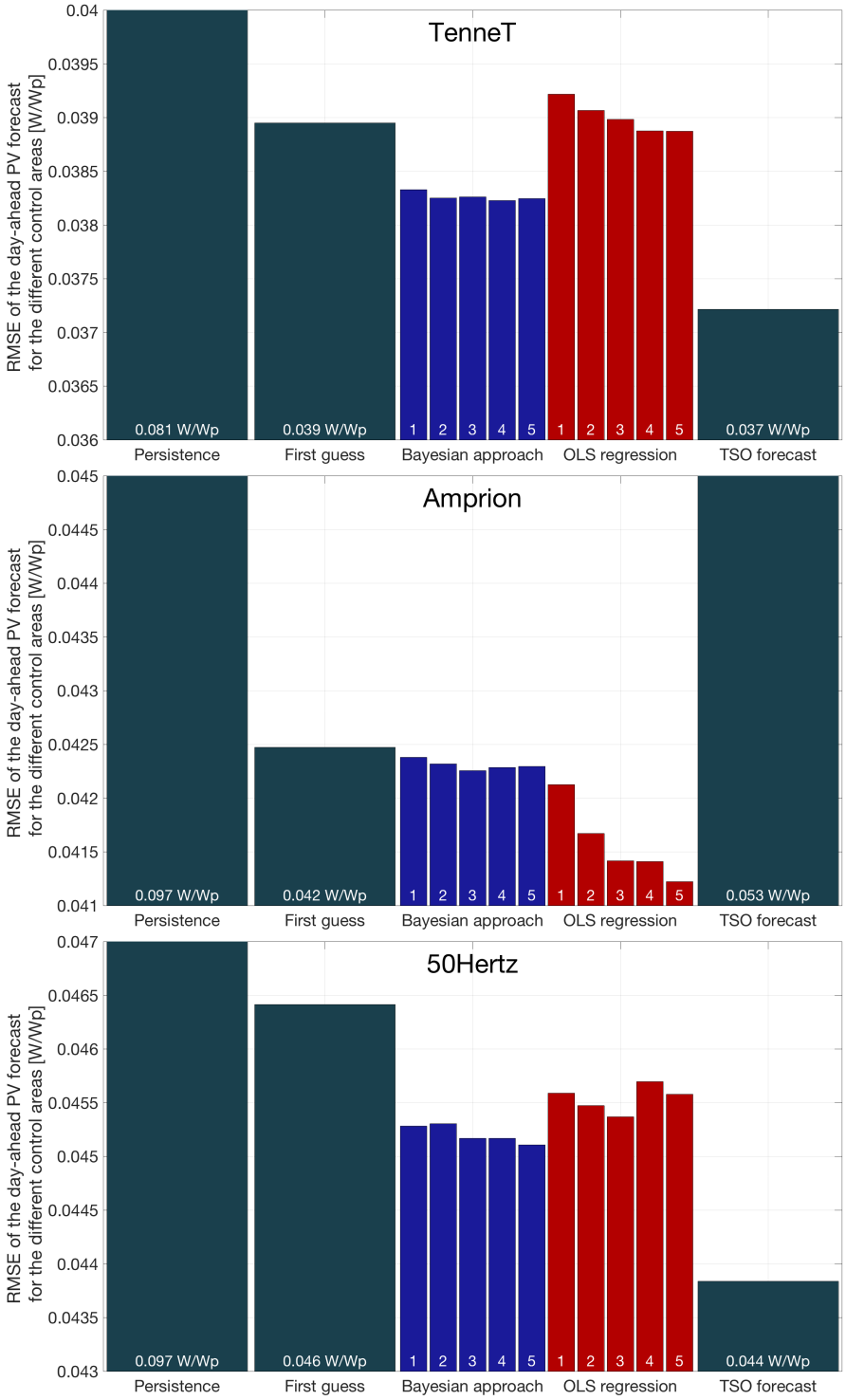


Figure 10: Comparison of the RMSE obtained with the different forecasts for TenneT (upper plot), Amprion (middle plot) and $50Hertz$ (lower plot)

OLS regression RMSE is better than initial guess RMSE. Thus, it is possible than the initial guess is sub-optimal but is balanced by the Bayesian inference, which ultimately yields a smaller RMSE than simple regression.

Our Bayesian method model output is finally compared to the TSO day-ahead forecast. TSO forecast RMSE values (excl. Amprion) are noticeably smaller than the alternative forecasts. TSO forecasts are the output of an optimized ensemble from several weather models; the ensemble is optimized daily by means of an adaptive method accounting for the actual value of the power production. In contrast, our simpler forecast only uses a single weather model that is optimized on a 12 month period. Given the substantial differences in required input data and adjustment techniques between the TSO and Bayesian forecasts, it is unsurprising that TSO forecasts have a better skill. Outperforming the TSO forecast is an unrealistic expectation with the proposed methodology. Instead, TSO forecasts are indicative expected performance from highly optimized forecasting approaches; this makes them a useful benchmark for the present evaluation. Figure 10 and Table 2 show that the difference between RMSE values obtained with the Bayesian method and TSO data decrease from 0.18 % (3.90 % - 3.72 %) with the initial guess to 0.10 % (3.82 % - 3.72 %) with the Bayesian method for Tennet; and from 0.26% (4.64 % - 4.38 %) to 0.14% (4.52 % - 4.38 %) at 50 Hertz. This result clearly demonstrates the potential of the proposed methodology with respect to the previous versions (Saint-Drenan et al., 2017). As the Bayesian method has performance close to TSO forecasting, we expect significant rewards from further methodological development. Potential improvement avenues are–but not limited to—consideration of an ensemble of NWP models, and incorporation of advanced MOS technique, such as exponential smoothing, Kalman Filter to track change of the fleet of PV systems, and mitigate NWP forecast error.

5. Discussion and conclusions

In this paper, a Bayesian method to parameterize a regional PV power forecasting model is proposed using only time series of regional PV power generation. This work addresses a need of forecast providers who must compete with transmission system operators (TSOs) estimates of the regional PV power generation without having access to critical information about the PV installations within the region.

The proposed approach is based on a linear regional PV model previously defined in literature (e.g. Saint-Drenan et al. (2017)). We made considerable effort to minimize the number of unknowns, especially when calibrating the reference orientation combinations and the selecting spatial constraints. Restricting of the number of predictors—or regularization by discretization—is interpreted as a projection method. We proposed a regularization that is based on a Bayesian problem formulation and describe a method to approximate the initial parameters as well as the covariance matrices required.

There is the possibility to consider different number of sub-regions in our model, whereby a larger region

is considered as the sum of all sub-regions. We found out that the sensitivity of the forecast performance on the number of sub-regions is very different for the 3 TSOs considered; while better performance was obtained from using more sub-regions for Amprion and TenneT, it is moderate to negative for 50 Hertz. This finding is probably a result of the actual PV system characteristics in the wider region as well as from the quality of the first guess and covariance matrices. Finally, it is clear that the model is prone to overtraining with an increasing number of degrees of freedom. Overtraining issues can be efficiently addressed by the proposed regularization technique. We illustrated this by exposing the coefficients as obtained by the OLS regression; they were frequently unrealistically high or even negative values. In parallel, coefficients from the Bayesian method were reasonable, positive and physically likely, resulting in a more robust parameterization.

Data from three German TSOs was used to evaluate forecasting performance. We find that, as long as the initial guess and covariance matrices are well-defined, the proposed approach offers significant added value with respect to persistence, an OLS regression, or to using only the initial guess. Interestingly, we identified that the Amprion TSO forecast was operating at significantly sub-optimal performance; for representation of our methodology, we excluded Amprion from our overall conclusions. A comparison of forecasts within TenneT and 50 Hertz regions showed that all approaches were better than persistence. The initial guess from our method was better than OLS regression with only < 4 sub-regions; however, OLS regression was better than the initial guess forecast at 50 Hertz. The Bayesian method was better than OLS regression and the initial guess. This is indicative of added value from Bayesian parametrization and from a parametrization approach.

With our method, the initial parameters and covariance matrices are inferred from a dataset containing metadata of numerous PV systems, however, information on installed PV systems in terms of installed capacity and spatial distribution are still required. Although our method has a reduced dependency on data availability than alternative methods, this initial parametrization is compensated by Bayesian parameter estimation; regardless, this overarching PV data is still needed. An interesting solution arises from the impressive work of collection, analysis and synthesis of PV system metadata statistics presented by Killinger et al. (2018), which constitutes exemplary research that can potentially improve the performances of regional forecasts in the future by removing the data dependency.

We trained the model parameters on a NWP weather forecast, selected because the targeted application is day-ahead regional PV power forecasting. It would be interesting to train the model parameters on alternative datasets, such as reanalysis or satellite data, to minimize the input error influence on the parameter estimation. Furthermore, integrate robust regression methods in the proposed approach to limit sensitivity to large input data errors is of interest.

In this work, we focused on the prediction of the regional PV power generation for the TSO but our approach can be used in any application where a regionally distributed fleet of PV systems needs to be modeled and information on the aggregated PV power production is available.

Finally, our approach could be extended to evaluate model parameters using different sources of information of different natures. The proposed framework indeed offers the possibility - with a minimal adaptation - to assimilate, in addition to aggregated power estimates, power measurements of individual PV systems or even aggregated power estimation of sub-regions (DSO estimates) to better assess the model parameters. This would allow getting more reliable information on the fleet of installed PV system but also better characterizing the PV systems by using more sub-regions.

Acknowledgements

The authors would like to warmly thank Amprion, TenneT and 50 Hertz for the pleasant cooperation and for the highly interesting discussions on the RES integration mechanisms.

This research was partly funded by the 'Bundesminsterium für Wirtschaft und Energie' and conducted within the project EWeLiNE (Erstellung innovativer Wetter-und Leistungsprognosemodelle für die Netzintegration wetterabhangiger Energieträger 0325500B).

J.M. Bright is funded by the Australian Renewable Energy Agency (ARENA, Research and Development Programme Funding G00854).

References

Amprion, 2019. Balancing group management. Accessed: 2019-02-27.
URL https://www.amprion.net/Energy-Market/Balancing-Groups/Balancing-Group-Management/

Boilley, A., Wald, L., 2015. Comparison between meteorological re-analyses from era-interim and merra and measurements of daily solar irradiation at surface. Renewable Energy 75, 135 – 143.
URL http://www.sciencedirect.com/science/article/pii/S0960148114006077

Bright, J. M., Killinger, S., Lingfors, D., Engerer, N. A., 2017. Improved satellite-derived PV power nowcasting using real-time power data from reference PV systems. Solar Energy.
URL http://www.sciencedirect.com/science/article/pii/S0038092X17309714

Colton, D., Kress, R., 1997. Inverse Acoustic and Electromagnetic Scattering Theory. Applied Mathematical Sciences. Springer Berlin Heidelberg.
URL https://books.google.fr/books?id=U5wemvlDnNwC

Crisan, A., Bain, D., 2009. Fundamentals of Stochastic Filtering, springer n Edition.

Crisan, D., Hambly, B., Zariphopoulou, T., 2014. Stochastic analysis and applications 2014: In honour of terry lyons.
URL http://dx.doi.org/10.1007/978-3-319-11292-3

Da Silva Fonseca, J. G., Oozeki, T., Ohtake, H., Shimose, K. I., Takashima, T., Ogimoto, K., 2014. Forecasting regional photovoltaic power generation - A comparison of strategies to obtain one-day-ahead data. Energy Procedia 57, 1337–1345.
URL http://dx.doi.org/10.1016/j.egypro.2014.10.124

Da Silva Fonseca Junior, J. G., Oozeki, T., Ohtake, H., ichi Shimose, K., Takashima, T., Ogimoto, K., 2014. Regional forecasts and smoothing effect of photovoltaic power generation in Japan: An approach with principal component analysis. Renewable Energy 68, 403–413.
URL http://dx.doi.org/10.1016/j.renene.2014.02.018

Engl, H., Hanke, M., Neubauer, A., 2000. Regularization of Inverse Problems. Mathematics and Its Applications. Springer Netherlands.
URL https://books.google.fr/books?id=VuEV-Gj1GZcC

Fonseca Junior, J. G. d. S., Oozeki, T., Ohtake, H., Takashima, T., Ogimoto, K., oct 2015. Regional forecasts of photovoltaic power generation according to different data availability scenarios: a study of four methods. Progress in Photovoltaics: Research and Applications 23 (10), 1203–1218.
URL http://doi.wiley.com/10.1002/pip.2528

Freitag, M., Potthast, R., 2013. Synergy of inverse problems and data assimilation techniques. Radon Series on Computational and Applied Mathematics. De Gruyter, Germany.

Hadamard, J., 1902. Sur les problèmes aux dérivées partielles et leur signification physique. Princeton Uni. Bull. (1902) 13, 49–52.

IEA, 2018. 2018: Snapshot of global photovoltaic markets: Report IEA PVPS T1-33:2018.
URL http://www.iea-pvps.org/fileadmin/dam/public/report/statistics/IEA_PVPS-A_Snapshot_of_Global_PV-1992-2017.pdf

Ineichen, P., 2014. Long term satellite global, beam and diffuse irradiance validation. Energy Procedia 48, 1586 – 1596, proceedings of the 2nd International Conference on Solar Heating and Cooling for Buildings and Industry (SHC 2013).
URL http://www.sciencedirect.com/science/article/pii/S187661021400441X

Killinger, S., 2017. A probabilistic approach to the estimation of regional photovoltaic power generation using meteorological data: Application of the Approach to the German Case. Ph.D. thesis, Karlsruher Institut für Technologie (KIT), Karlsruhe, Germany.
URL http://publica.fraunhofer.de/eprints/urn_nbn_de_0011-n-4942211.pdf

Killinger, S., Lingfors, D., Saint-Drenan, Y.-M., Moraitis, P., van Sark, W., Taylor, J., Engerer, N., Bright, J., Oct. 2018. On the search for representative characteristics of PV systems: Data collection and analysis of PV system azimuth, tilt, capacity, yield and shading. Solar Energy 173, 1087 – 1106.
URL https://hal.archives-ouvertes.fr/hal-01882680

Lorenz, E., Hurka, J., Karampela, G., Heinemann, D., Beyer, H., Schneider, M., 2008. Qualified Forecast of ensemble power production by spatially dispersed grid-connected PV systems. 23rd European Photovoltaic Solar Energy Conferenc (January), 3285–3291.
URL http://tasks.iea-shc.org/publications/downloads/23rd{_}EU{_}PVSEC{_}5A0.8.6{_}lorenz.pdf

Lorenz, E., Scheidsteger, T., Hurka, J., Heinemann, D., Kurz, C., nov 2011. Regional PV power prediction for improved grid integration. Progress in Photovoltaics: Research and Applications 19 (7), 757–771.
URL http://doi.wiley.com/10.1002/pip.1033

Nakamura, G., Potthast, R., 2015. Inverse Modeling: An Introduction to the Theory and Methods of Inverse Problems and Data Assimilation. IOP expanding physics. Institute of Physics Publishing.
URL https://books.google.fr/books?id=toR_jgEACAAJ

Nuno Martinez, E., Koivisto, M. J., Cutululis, N. A., Sorensen, P., 2018. On the simulation of aggregated solar PV forecast errors. IEEE Transactions on Sustainable Energy, 1–10.

Saint-Drenan, Y., Good, G., Braun, M., 2017. A probabilistic approach to the estimation of regional photovoltaic power production. Solar Energy 147, 257–276.

Saint-Drenan, Y.-M., 2015. A probabilistic approach to the estimation of regional photovoltaic power generation using meteorological data: Application of the Approach to the German Case. Ph.D. thesis, University of Kassel, Kassel, Germany.
URL https://kobra.bibliothek.uni-kassel.de/bitstream/urn:nbn:de:hebis:34-2016090550868/3/DissertationYMSaintDrenan.pdf

Schierenbeck, S., Graeber, D., 2010. Ein distanzbasiertes Hochrechnungsverfahren für die Einspeisung aus Photovoltaik. Energiewirtschaftliche tagesfragen 60 (12), 60–64.
URL http://www.transnetbw.de/downloads/kennzahlen/erneuerbare-energien/2010-fachartikel-graeber-semmig-schierenbeck-weber-pdf

Schubert, G., 2012. Modeling hourly electricity generation from PV and wind plants in Europe. 9th International Conference on the European Energy Market, EEM 12, 1–7.

Shaker, H., Zareipour, H., Wood, D., 2015. A Data-Driven Approach for Estimating the Power Generation of Invisible Solar Sites. IEEE Transactions on Smart Grid PP (99), 1–11.

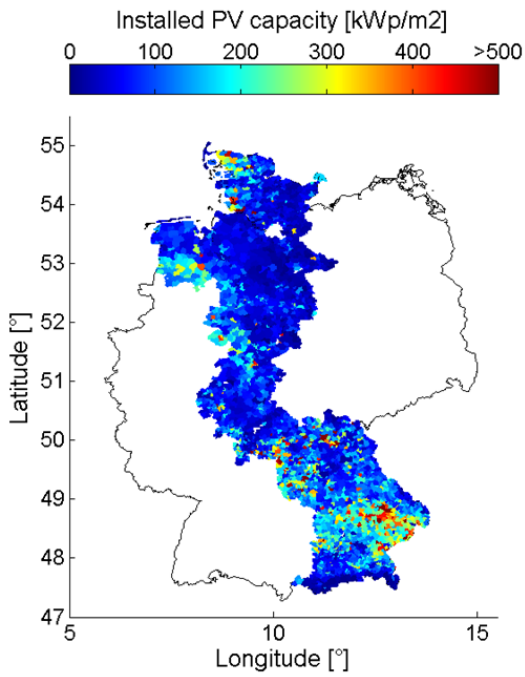
Tikhonov, A. N., 1963. Solution of Incorrectly Formulated Problems and the Regularisation Method.pdf. Soviet Mathematics Doklady 4 (4), 1035–1038.

Wilks, D. S., 2011. Statistical methods in the atmospheric sciences. Elsevier Academic Press, Amsterdam; Boston.
URL https://www.amazon.com/Statistical-Atmospheric-Sciences-International-Geophysics/dp/0123850223/ref=pd_bxgy_14_img_3?_encoding=UTF8&psc=1&refRID=ESPQQ0R2PB1TP1VJSGCZ

Yang, D., Quan, H., Disfani, V. R., Rodríguez-Gallegos, C. D., 2017. Reconciling solar forecasts: Temporal hierarchy. Solar Energy 158, 332–346.

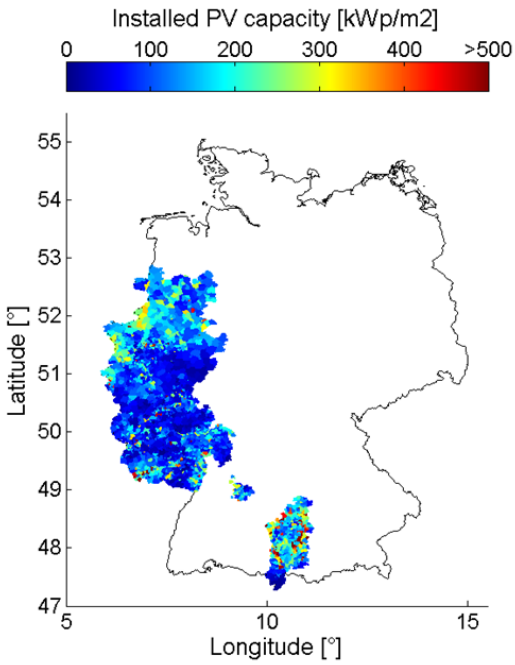
Appendix A. Description of the control area of the German TSOs

Description of the control area of TenneT



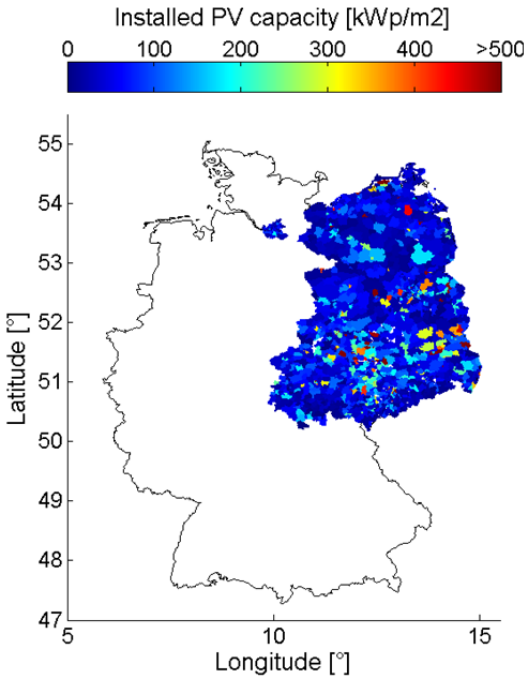
Main characteristics of the TenneT control area (Update time: 05/2016)		
Area		140.000 km^2
Time of the last input		*12-May-2016*
Number of plants	0-30kW$_p$	595 415 (88.7 %)
	30-100 kW$_p$	63 347 (9.4 %)
	100-1000 kW$_p$	11 346 (1.7 %)
	>1000 kW$_p$	1 143 (0.2 %)
	Total	**671 251 (100 %)**
Installed capacity	0-30kW$_p$	6 569.2 MW$_p$ (42.8 %)
	30-100 kW$_p$	3 075.3 MW$_p$ (20.0 %)
	100-1000 kW$_p$	2 591.7 MW$_p$ (16.9 %)
	>1000 kW$_p$	3 115.9 MW$_p$ (20.3%)
	Total	**15 352.0 MW$_p$ (100 %)**

Description of the control area of Amprion



Main characteristics of the Amprion control area (Update time: 05/2016)		
Area		73.100 km^2
Time of the last input		*03-May-2016*
Number of plants	0-30kW$_p$	411 271 (89.3 %)
	30-100 kW$_p$	40 387 (8.8 %)
	100-1000 kW$_p$	8 623 (1.9 %)
	>1000 kW$_p$	501 (0.1 %)
	Total	**460 782 (100 %)**
Installed capacity	0-30kW$_p$	4 004.5 MW$_p$ (43.1 %)
	30-100 kW$_p$	2 039.2 MW$_p$ (21.9 %)
	100-1000 kW$_p$	1 927.7 MW$_p$ (20.7 %)
	>1000 kW$_p$	1 329.2 MW$_p$ (14.3 %)
	Total	**9 300.6 MW$_p$ (100 %)**

Description of the control area of 50 Hertz



Main characteristics of the 50 Hertz control area *(Update time: 05/2016)*		
Area		109.360 km^2
Time of the last input		20-Apr-2016
Number of plants	0-30kW$_p$	118 296 (85.3%)
	30-100 kW$_p$	12 910 (9.3 %)
	100-1000 kW$_p$	5923 (4.3%)
	>1000 kW$_p$	1565 (1.1%)
	Total	**138 694 (100%)**
Installed capacity	0-30kW$_p$	983.2 MW$_p$ (11.0%)
	30-100 kW$_p$	626.5 MW$_p$ (7.0 %)
	100-1000 kW$_p$	1 777.5 MW$_p$ (19.9 %)
	>1000 kW$_p$	5 533.2 MW$_p$ (62.0 %)
	Total	**8920.4 MW$_p$ (100 %)**

638

a) TenneT control area

Number of values: 12 473

Average power: 0,2091 W/W_p

	Persistence	First guess	Bayesian method min / max	OLS regression min / max	TSO forecast
correlation [-]	0,8815	0,9724	0.9737 / 0.9738	0.9723 / 0.9729	0,9806
bias [W/W_p]	-0,0008	-0,0038	-0.0059 / -0.0057	-0.006 / -0.0053	0,0131
std [W/W_p]	0,0808	0,0388	0.0378 / 0.0379	0.0384 / 0.0388	0,0348
MAE [W/W_p]	0,0538	0,0263	0.0258 / 0.0259	0.0264 / 0.0266	0,0249
RMSE [W/W_p]	0,0808	0,0390	0.0382 / 0.0383	0.0389 / 0.0392	0,0372

b) Amprion control area

Number of values: 12 451

Average power: 0,2307 W/W_p

	Persistence	First guess	Bayesian method min / max	OLS regression min / max	TSO forecast
correlation [-]	0,8601	0,9731	0,9730 / 0,9732	0,9735 / 0,9746	0,9638
bias [W/W_p]	0,0001	-0,0034	-0,0015 / -0,0013	-0,0032 / -0,0028	0,0013
std[W/W_p]	0,0972	0,0423	0,0422 / 0,0424	0,0411 / 0,0420	0,0527
MAE [W/W_p]	0,0637	0,0286	0,0284 / 0,0285	0,0280 / 0,0287	0,0360
RMSE [W/W_p]	0,0972	0,0425	0,0423 / 0,0424	0,0412 / 0,0421	0,0527

c) 50 Hertz control area

Number of values: 12 111

Average power: 0,2380 W/W_p

	Persistence	First guess	Bayesian method min / max	OLS regression min / max	TSO forecast
correlation [-]	0,8710	0,9712	0,9720 / 0,9722	0,9716 / 0,9720	0,9736
bias [W/W_p]	0,0022	0,0018	-0,0046 / -0,0045	-0,0048 / -0,0043	-0,0011
std [W/W_p]	0,0973	0,0464	0,0449 / 0,0451	0,0451 / 0,0454	0,0438
MAE [W/W_p]	0,0646	0,0302	0,0299 / 0,0300	0,0299 / 0,0302	0,0285
RMSE [W/W_p]	0,0973	0,0464	0,0451 / 0,0453	0,0454 / 0,0457	0,0438

Table 2: Different error metrics obtained with the various forecast approaches for the control areas of a) TenneT, b) Amprion and c) 50 Hertz