Probabilistic prediction of solar power supply to distribution networks, using forecasts of global horizontal irradiation

F. von Loeper[a,*], P. Schaumann[a], M. de Langlard[a], R. Hess[b], R. Bäsmann[c], V. Schmidt[a]

[a] *University Ulm, Institute of Stochastics, Helmholtzstraße 18, 89081 Ulm, Germany*
[b] *Deutscher Wetterdienst, Frankfurter Straße 135, 63067 Offenbach, Germany*
[c] *N-ERGIE Netz GmbH, Sandreuthstraße 21, 90441 Nürnberg, Germany*

Abstract

This paper presents a mathematical model for the prediction of the probabilities of reverse power flow exceeding predefined critical thresholds at feed-in points of a distribution network. The parametric prediction model is based on hourly forecasts of global horizontal irradiation and uses copulas, a tool for modeling the joint probability distribution of two or more strongly correlated random variables with non-Gaussian (marginal) distributions. The model is used for determining the joint distribution of forecasts of global horizontal irradiation and measured solar power supply at given feed-in points, where respective sample datasets were provided by Deutscher Wetterdienst and the N-ERGIE Netz GmbH. It is shown that the fitted model replicates important characteristics of the data such as the corresponding marginal densities. The validation results highlight strong performance of the proposed model. The copula-based model enables to predict the distribution of solar power supply conditioned on the forecasts of global horizontal irradiation, thus anticipating great fluctuations in the distribution network.

Keywords: Probabilistic prediction model, Global horizontal irradiation, Solar power supply, Mixed beta distribution, Archimedean copula

---

*Corresponding author
Email address: `freimut.von-loeper@uni-ulm.de` (F. von Loeper)

Preprint submitted to Elsevier *March 31, 2020*

1. Introduction

In the recent decade, the global annually installed capacity of solar power increased rapidly, reaching 98 Gigawatts in 2017 alone. Compared to other power generation sources connected to electricity networks, solar power has the greatest capacity installed in 2017, followed by wind power with 52 Gigawatts, gas power with 38 Gigawatts, coal power with 35 Gigawatts and various other sources adding up to 37 Gigawatts. Although the global capacity of solar power installed in 2017 already exceeded most expectations, solar analysts predict even further increase of the annually installed capacity for the future. Thus, it comes as no surprise that the worldwide installed capacity of solar power is estimated to exceed 1 Terrawatt by 2022 (SolarPower Europe, 2017).

The increase in solar penetration causes greater fluctuations in the power supply, which might result in increasing overloading problems and voltage violations (Karimi et al., 2016). To tackle the upcoming challenges for distribution network operators, suitable computer-based models are developed. Since solar power is a variable power source, better forecasts of solar power supply are useful to predict oversupply events and reduce the curtailment (Bird et al., 2014). Moreover, planning maintenance becomes easier and bids in electricity markets can be better estimated. For a deeper understanding of the economical value of forecasting solar power supply we refer to Antonanzas et al. (2016).

Most prediction models of solar power supply rely on weather forecasts as input (Antonanzas et al., 2016). Usually, these weather forecasts are the result of a complex modeling process starting with numerical weather prediction models based on data assimilation that use all kinds of meteorological observations in real time including e.g. temperature, wind, pressure, humidity, as well as geospatial data as input information (Coiffier, 2011). Based on this data, the task of numerical weather prediction models such as the high resolution version COSMO-DE of the Consortium for Small-scale Modeling (COSMO) run by Deutscher Wetterdienst (DWD) is to solve differential equations for modeling the spatio-temporal evolution of meteorological variables and simulate physical processes (Baldauf et al., 2011). However, numerical weather prediction models are subject to systematic errors and not all weather phenomena are simulated and need to be interpreted therefore. For these reasons, statistical post-processing methods are applied. For example, DWD runs Model Output Statistics (MOS) techniques (Hess et al., 2015), which are based on multiple linear and logistic regression models (Jobson, 1991). The regression models fit historical data of the direct output of numerical models to corresponding observations from synoptic stations. In operational use, the fitted regression models calibrate and interpret the output of numerical weather predictions resulting in deterministic and probabilistic forecasts for various meteorological variables (Heinemann et al., 2006).

Since energy conversion by solar plants mainly depends on direct and diffuse radiation, our prediction model is based on deterministic and statistically post-processed forecasts of global horizontal irradiation (GHI). Other meteorological variables such as ambient temperature, wind velocity, humidity and dust, may also influence the energy conversion and have been used in the literature (Kaldellis et al., 2014; Mekhilef et al., 2012). Moreover, probabilistic models which take into account the spatio-temporal correlation between these variables were also proposed. Almeida et al. (2015) developed a non-parameteric quantile regression forest model which takes as training input temperature, wind speed, wind direction, humidity, sea level pressure and cloud cover at different levels. Bessa et al. (2015) proposed a vector auto-regressive model with time series information collected at different locations on a smart grid as input. In Zhang et al. (2016) a Gaussian conditional random field model was applied, where historical forecasts and solar power measurements were considered at many solar sites. Huang and Perry (2016) estimated prediction intervals based on a k-nearest neighbor regression and added to deterministic forecasts computed by gradient boosting, where weather variables such as solar radiation, temperature, cloud ice water content, wind speed were considered. Solar power measurements at adjacent solar farms were included as explanatory regression variables. In contrast we apply a copula model which only considers the GHI forecasts, the most important explanatory variable for energy conversion. This simplifies the fitting procedure of the model due to a smaller number of parameters, while still capturing the correlation structure between weather conditions and solar power supply.

Copulas are a mathematical tool to model the joint distribution of two or more random variables. In the context of renewable energies, they were first applied for the probabilistic prediction of wind power generation, see e.g. Papaefthymiou and Kurowicka (2009); Wang et al. (2014); Lu et al. (2014). Recently, copula models have also been used for statistical analysis of data on solar power generation. For example, Golestaneh et al. (2016a,b) applied quantile regression to non-parametrically compute conditional marginal densities of solar power supply for neighboring solar plants, given numerical weather prediction forecasts. Then, in a next step, multivariate Gaussian copulas were used in Golestaneh et al. (2016a) to determine the joint conditional distribution of solar power supply at neighboring plants with the previously computed non-parametric conditional marginal densities. Golestaneh and Gooi (2017) compared Gaussian copulas with multivariate R-vine copulas. However, in both papers copulas were merely used to model the spatial relationship between solar power supply at neighboring plants. More recently, Panamtash et al. (2020) proposed a similar copula-based model, where bivariate copulas are applied to improve prior probabilistic forecasting done by traditional forecasting methods such as multiple linear regression, artificial neural networks, gradient boosting, random forests and autoregressive integrated moving average. The results showed that copulas capture the joint probability distribution of solar power and temperature effectively. In Panamtash et al. (2020), ambient temperature has been considered as input variable, while we focus on the correlation between GHI forecasts and solar power supply.

Given a weather forecast, our proposed model computes conditional probabilities of reverse power flow exceeding predefined critical thresholds at feed-in points of a distribution network. To implement the prediction model, the first step is to fit univariate (so-called marginal) probability distributions using historical data of hourly GHI forecasts and hourly averages of measured solar power supply. The second step is to model the joint probability distribution of GHI forecast and solar power supply by applying copula theory (Durante and Sempi, 2015; Nelsen, 2006; Joe, 2014). Finally, the third step is to compute the conditional probability distribution of solar power supply based on the fitted joint and marginal distributions. Taking a real-time GHI forecast as input, a probabilistic prediction of solar power supply for the same time horizon as the weather forecast can be computed by the prediction model.

The rest of this paper is organized as follows. In Section 2, the data used in this paper is described, including its pre-processing and analysis. The model and its fitting procedure is explained in Section 3. The fitted model characteristics and the validation of the prediction model are discussed in Section 4, where also the performance of the proposed model is compared with that of the quantile regression technique, one of the most frequently used probabilistic prediction method (see, for instance, Bacher et al. (2009); Zamo et al. (2014); Massidda and Marrocu (2018); Lauret et al. (2017); Golestaneh et al. (2016b); Alessandrini et al. (2015)). Finally, Section 5 concludes.

2. Data

The modeling approach for the prediction of solar energy supply, proposed in the present paper, is a parametric probabilistic model which is based on copulas (Durante and Sempi, 2015; Nelsen, 2006; Joe, 2014). Compared to conventional photovoltaic performance models, a probabilistic model is more flexible, but needs historical data as input (Antonanzas et al., 2016). In particular, the general modeling idea is not concerned with physical attributes of the datasets, e.g. the locations of the measurement points. Our probabilistic modeling approach and its application are illustrated by using suitable sample datasets provided by DWD and the N-ERGIE Netz GmbH (NNG). For calibration and validation of our model, the time frame covering the months May, June and July of the years 2015 till 2017 is considered, resulting in 273 days.

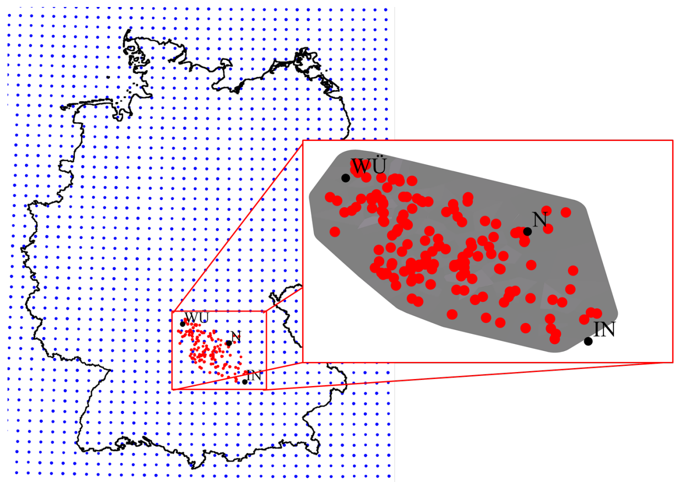


Figure 1: Forecast grid of DWD (blue) and feed-in points of NNG (red). The (appropriately dilated) convex hull (grey) of all feed-in points in the zoom-in of the red box is the part of Germany used for visualizing our results. To illustrate the geographical location of this area, the cities of Würzburg (WÜ), Nürnberg (N) and Ingolstadt (IN) are depicted (black).

2.1. Description of data

The fitting of our model is based on two datasets, namely GHI forecasts provided by DWD and measured solar power supply provided by NNG.

The first dataset consists of hourly GHI forecasts (in kJ/m^2), statistically interpreted based on synoptic observations and numerical forecasts of COSMO-DE-EPS, the ensemble system of COSMO-DE at DWD. The forecasts are issued every three hours with forecast lead times up to 19 hours. For the time frame mentioned above, the forecasts are available on a $20\,\mathrm{km} \times 20\,\mathrm{km}$ grid covering Germany and parts of neighboring countries, see Figure 1. However, there is no GHI forecast generated for grid points and forecasts times with a local solar elevation angle of less than 5 degrees at the beginning or end of the forecast hour.

The second dataset, provided by NNG, records the amount of electricity, which was generated by solar plants, supplied to the distribution network and measured at its feed-in points. These amounts of solar power supply are 15-minute average values. Each of the 168 feed-in points, considered in this paper, is connected with at least one solar plant with nominal capacity of the connected solar plants ranging from 0.25 up to 10 Megawatts.

In Figure 1 the considered feed-in points of NNG are visualized. The corresponding supply area covers ca. 8000 km^2.

2.2. Data preprocessing

At first temporal and spatial compatibility between the datasets has to be established. Temporal compatibility can be easily realized by calculating the averages of solar power supply for each hour. For spatial compatibility we consider two different hierarchy levels in the distribution network. On the one hand we are interested in the GHI forecasts and amounts of solar power supply at the feed-in points, on the other hand we want to apply our model to communities, i.e. sets of neighboring feed-in points, as well. Therefore, we match GHI forecasts and amounts of solar power supply to feed-in points and communities, see Section 2.2.1 and 2.2.2.

2.2.1. Spatial compatibility for feed-in points

In this section we consider the problem to match the amount of solar energy measured at every feed-in point to a single GHI forecast. For simplicity, the locations of solar plants are assumed to coincide with the locations of their feed-in points. Since the hourly averages of forecasted GHI are practically the same on such small spatial scales, the error introduced by this assumption is negligible. The GHI forecast at a certain feed-in point has been estimated by interpolating the GHI forecast at the grid points of the 20 km $\times$ 20 km grid, see Figure 1. This is done by bilinear interpolation, see Hämmerlin and Hoffmann (2012).

2.2.2. Spatial compatibility for communities

In a next step, the amounts of solar power supply and the GHI forecasts need to be matched to communities. Therefore, we approximate the solar power supply generated in a community by summing the solar power supply measured at the feed-in points over all feed-in points within the community. It is assumed that there are no transmission losses, but alternatively the losses can be computed either using explicit formulas or statistical estimations, see Dickert et al. (2009), Council of European Energy Regulators (2017).

To compute the GHI forecasts for the corresponding communities, we use the interpolated GHI forecasts at the feed-in points in a community as mentioned in Section 2.2.1. By averaging the GHI forecasts over all feed-in points in a community, we get the GHI forecast of the community.

2.2.3. Selection of data

Due to maintenance, repair work and risk of overloading, some solar plants might have to be shut down for certain time periods. Since these actions are not directly related with weather phenomena, corresponding time periods are excluded by removing solar power supply being equal to zero from data. The GHI forecasts for those locations and forecast times are also removed from data, leaving only data pairs with matching time stamps.

Furthermore, the performance of solar plants is strongly influenced by many factors apart from meteorological variables, e.g. nominal capacity, tilt angle and composition of photovoltaic units. Most of these factors are constant over long time periods, but their influence might largely depend on the time of day as it is the case with the tilt angle. A simple way to remove such effects from data is to consider each hour of the day and feed-in point separately.

Lastly, by matching both data sets, all hours which have no measurement of solar power supply or GHI forecast are removed. This includes all night hours, see Section 2.1.

2.2.4. Rescaling the data

For an easier comparison of the datasets with different scales, the interpolated GHI forecasts and measured amounts of solar power are locally normalized. More precisely, the datasets are rescaled for a certain feed-in point, or community, by applying the transformation

$$\phi_{a,b}(x) = (1 - 2c)(x - a)/(b - a) + c, \tag{1}$$

where $c > 0$ is close to zero. If a is the minimum and b the maximum of the dataset under consideration, then, ϕ maps onto the interval $[c, 1-c]$. For $c > 0$ close to zero, $[c, 1-c]$ approximates the open interval $(0, 1)$. Thus, we use $c = 0.001$ in this paper.

2.3. Empirical data analysis

The rescaled and interpolated data of solar power supply and GHI forecast for feed-in points are analyzed in order to highlight their strong correlation and their spatial disparities. We consider the time frame May, June and July of the years 2015 till 2017 (11-12 UTC), denoted by T, and GHI forecasts with forecast lead time of one hour. The method, described in Section 2.2, is applied for each feed-in point and dataset separately, where the rescaling parameters a and b, in Section 2.2.4, are set to the minimum and maximum of each dataset in T.

Therefore, the local *empirical correlation coefficient* of GHI forecasts and amounts of solar power supply at each feed-in point ℓ, denoted by $\rho(\ell)$, is computed. Note that the empirical correlation coefficient $\rho(\ell)$ is defined as

$$\rho(\ell) = \frac{\sum_{t\in T}(r(\ell,t) - \overline{r}(\ell))(s(\ell,t) - \overline{s}(\ell))}{\sqrt{\sum_{t\in T}(r(\ell,t) - \overline{r}(\ell))^2 \sum_{t\in T}(s(\ell,t) - \overline{s}(\ell))^2}}, \tag{2}$$

where $r(\ell,t)$ and $s(\ell,t)$ are the preprocessed GHI forecast and solar power supply for forecast time t, respectively, and $\overline{r}(\ell)$ and $\overline{s}(\ell)$ are the corresponding time averages.

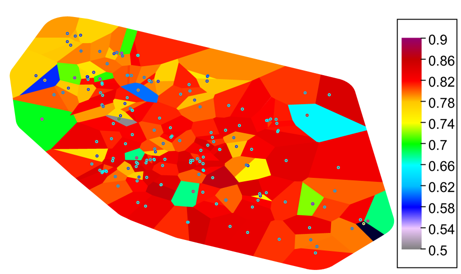
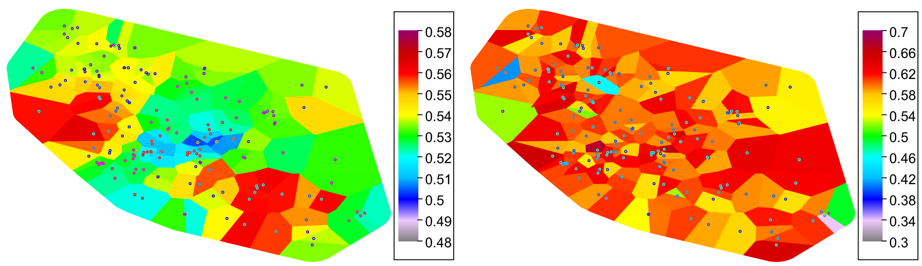


Figure 2: Empirical correlation coefficients of preprocessed GHI forecasts and preprocessed amounts of solar power supply for each feed-in point visualized by Voronoi tessellation.

In Figure 2, the results are visualized, which we obtained for the local empirical correlation coefficients, where the Region of Interest (ROI), i.e., the (appropriately dilated) convex hull of all feed-in points, is decomposed into the Voronoi tessellation generated by the feed-in points. The value computed for a feed-in point is assigned to the entire Voronoi cell of this point.



(a) GHI forecast

(b) Solar power supply

Figure 3: Local means over 273 days for each feed-in point visualized by Voronoi tessellation.

In general, Figure 2 shows rather high empirical correlation coefficients for all feed-in points, but also the existence of significant differences between feed-in points. By depicting the local means of the preprocessed datasets in Figure 3, which are obtained by averaging over the 273 days considered in this paper, this observation becomes even more evident. Indeed, the map of local means of the GHI forecasts does not show the same kind of pattern as the one of the solar power supply, see Figure 3. Thus, there have to be other factors except GHI forecasts, e.g. physical characteristics of the solar plants connected with the feed-in points, influencing the local means of solar power supply shown in Figure 3.

As a conclusion of this empirical analysis, our modeling approach has to consider the interdependence of GHI forecasts and solar power supply, see Figure 2. Moreover, Figure 3 indicates that the parameters of the probability distributions considered in this paper should be determined for each feed-in point separately to take into account their spatial variability.

3. Copula-based model for the prediction of solar power supply

In many fields, where risk must be managed, probabilistic predictions are preferred as they allow to quantify the uncertainty. In our case, given some probability estimation for the occurrence of a critical feed-in event, distribution network operators might make their decisions individually based on how much risk they want to take. Further advantages of probabilistic forecasts are discussed, e.g., in Antonanzas et al. (2016).

3.1. Modeling approach

For simplicity, we consider preprocessed solar power supply and preprocessed GHI forecast at a certain location in the distribution network for a single time of day and forecast lead time. Both are interpreted as realizations of some random variables R and S, which are strongly correlated as Figure 2 in Section 2.3 depicts. The support of R and S is the interval $[0,1]$, because of the rescaling transformation given in Eq. (1).

Given a GHI forecast r the conditional probability of solar power supply exceeding a certain threshold $v \in [0,1]$ can be written in the following form:

$$P(S \geq v \mid R = r) = \int_v^1 f_{S|R}(s \mid r)\mathrm{d}s \tag{3}$$

$$= \int_v^1 \frac{f_{(R,S)}(r,s)}{f_R(r)}\mathrm{d}s, \tag{4}$$

where $f_{S|R}$ is the conditional density of the random variable S given R, $f_{(R,S)}$ the joint density of the random variables S and R, and f_R the marginal density of R. Thus, our task lies in modeling the marginal and joint distributions of S and R.

The proposed method can be decomposed in the following two steps:

1. derive a parametric form of the marginal densities f_R and f_S of the random variables R and S ;
2. use a copula and the marginal densities f_R and f_S to model the joint density function $f_{(R,S)}$.

Then, the conditional level-crossing probabilities $P(S \geq v \mid R = r)$ can be computed using Eq. (4). In the following, these two steps are further explained in detail.

3.1.1. Model for the marginal densities

Vale (2015) investigated several types of parametric distributions for solar irradiation in Lisbon with respect to the day time and month, where it is concluded that the mixed beta distribution is the best fit for most months of the year. However, the difference in the considered time frame and location might influence the quality of fits for the tested distribution types.

In Section 4.1.1, we demonstrate that the mixed beta distribution is indeed suitable for GHI forecasts and also for solar power supply. Therefore, we choose the models of the marginal densities f_S and f_R to be a mixture of beta densities. The family of beta distributions allows for various shapes of probability density functions.

Given a mixing parameter $q \in (0, 1)$ and two densities of beta distributions $f_i : \mathbb{R} \to [0, \infty)$ with $i \in \{1, 2\}$, the probability density $f_X : \mathbb{R} \to [0, \infty)$ of the *mixed beta distribution* of a random variable X is given by

$$f_X(x) = qf_1(x) + (1 - q)f_2(x) \tag{5}$$

for all $x \in \mathbb{R}$. For $i \in \{1, 2\}$, the *beta density* f_i has two shape parameters $a_i > 0$ and $b_i > 0$, and is defined by

$$f_i(x) = \frac{\Gamma(a_i + b_i)}{\Gamma(a_i)\Gamma(b_i)} x^{a_i - 1}(1 - x)^{b_i - 1} \tag{6}$$

for $x \in (0, 1)$ and $f_i(x) = 0$ otherwise, where Γ denotes the gamma function. Then, the marginal densities f_R and f_S take the form of the mixed density given in Eq. (5) and are specified by five parameters each. The determination of these parameters is explained in Section 3.2.

3.1.2. The copula model

If R and S were independent random variables, the bivariate density function $f_{(R,S)}$ would turn out to be the product of the univariate densities f_R and f_S.

But, in fact, the random variables R and S are strongly correlated as depicted in Figure 2. Therefore, the design of a parametric joint density is more complex.

To calculate the joint density with the non-Gaussian marginal densities fitted in Section 3.2, a parametric modeling approach based on Sklar's theorem is applied (Durante and Sempi, 2015). This fundamental result of copula theory allows us to represent the bivariate joint distribution function of two random variables by superposing a copula function upon the marginal distribution functions. Note that a *copula* is defined as the joint cumulative distribution function $C : [0,1] \times [0,1] \to [0,1]$ of a two-dimensional random vector (U, V) with components U and V uniformly distributed on $[0,1]$, see Nelsen (2006) for further details.

Let (R, S) be the two-dimensional random vector consisting of the random variables R and S introduced above with joint cumulative distribution function $F_{(R,S)} : \mathbb{R}^2 \to [0,1]$ and marginal distribution functions F_R and F_S. Then, *Sklar's theorem* says that a copula function $C : [0,1] \times [0,1] \to [0,1]$ exists such that

$$F_{(R,S)}(r, s) = C(F_R(r), F_S(s)) \tag{7}$$

for all $r, s \in \mathbb{R}$. Note that Eq. (7) can be written in the following differential form:

$$f_{(R,S)}(r, s) = f_R(r) \cdot f_S(s) \cdot c(F_R(r), F_S(s)), \tag{8}$$

where $f_{(R,S)}$, f_R, f_S and c are the densities corresponding to the cumulative distribution functions $F_{(R,S)}$, F_R, F_S and C. Using (8), the conditional density function $f_{S|R}(s \mid r)$ of solar power supply given a GHI forecast r can be computed by

$$f_{S|R}(s \mid r) = \frac{f_{(R,S)}(r, s)}{f_R(r)} = f_S(s) \cdot c(F_R(r), F_S(s)). \tag{9}$$

Thus, to determine the conditional level-crossing probability considered in (4), we estimate the marginal densities f_S and f_R, which leads to estimates of the corresponding distribution functions F_S and F_R, and the copula density c in Eq. (9). For the estimation of c we apply *Archimedean copulas*. Archimedean copulas are a commonly considered class of copulas which can be given by analytical formulas and are therefore especially easy to handle.

A function $g : [0,1] \to [0,\infty]$ is called an *Archimedean generator* if g is continuous, strictly decreasing and solves $g(1) = 0$. The pseudo-inverse $g^{[-1]}$ of an Archimedean generator g is an extension of the inverse function $g^{(-1)}$ defined as

$$g^{[-1]}(t) = \begin{cases} g^{(-1)}(t), & \text{if } 0 \leq t \leq g(0), \\ 0, & \text{if } g(0) < t \leq \infty. \end{cases} \tag{10}$$

The Archimedean copula generated by g is then given by

$$C(u,v) = g^{[-1]}(g(u) + g(v)) \tag{11}$$

for $u, v \in [0,1]$.

In this paper, we focus on four parametric types of Archimedean copulas, see Table 1, each of them having a single parameter $\theta \in \mathbb{R}$ to be fitted.

Type	Archimedean generator	Parameter
Joe	$g_\theta(t) = -\log(1-(1-t)^\theta)$	$\theta \in [1,\infty)$
Frank	$g_\theta(t) = (-\log(\frac{\exp(-\theta t-1)}{\exp(-\theta)-1}))^\theta$	$\theta \in \mathbb{R}\backslash\{0\}$
Clayton	$g_\theta(t) = \frac{1}{\theta}(t^{-\theta}-1)$	$\theta \in [-1,\infty)\backslash\{0\}$
Gumbel	$g_\theta(t) = (-\log(t))^\theta$	$\theta \in [1,\infty)$

Table 1: Types of Archimedean copulas

3.2. Model fitting procedure

In this section, we describe the procedure to find the best model parameters of the marginal densities f_S and f_R, which are modeled by mixtures of beta densities, see Eq. (5). Next, we detail the method to search for the copula type and the copula parameter θ, which gives us the best fit to the data. Both fitting procedures are based on the *maximum likelihood estimation* (MLE) principle, see Wilks (2011).

3.2.1. Fitting marginal densities

There are five parameters to be determined for each marginal density f_R and f_S: the mixing parameter q, the shape parameters a_1 and b_1 of the first beta density f_1, and the shape parameters a_2 and b_2 of the second beta density f_2, see Eq. (5).

The main idea is to consider the maximum of a certain product of likelihood functions. Note that the likelihood function is defined, in the case of the random variable R, by

$$L(\beta \mid r) = f_R(r), \tag{12}$$

where $\beta = (q, a_1, b_1, a_2, b_2)$ is the set of parameters and r is a GHI forecast. If we consider the dataset $r_1, \ldots, r_n$ and assume that the observations $r_1, \ldots, r_n$ are independently sampled realizations of the random variable R (n is the total number of observations), then the MLE consists in maximizing the function

$$\begin{aligned} L(\beta \mid r_1, \ldots, r_n) &= \prod_{i=1}^{n} L(\beta \mid r_i) && (13)\\ &= \prod_{i=1}^{n} f_R(r_i) && (14)\\ &= \prod_{i=1}^{n} q f_1(r_i) + (1-q) f_2(r_i), && (15) \end{aligned}$$

where the densities f_1 and f_2 depend on the parameters a_1 and b_1, respectively a_2 and b_2. The MLE expresses the fact that the realized observations occur with the highest possible probability. In general, it is more convenient to take the logarithmic form of the expression given in (15) in order to deal with a sum instead of product operations. Then, the best parameters q, a_1, b_1, a_2 and b_2 for the marginal density of GHI forecasts are the solution of the maximization problem

$$\underset{q, a_1, b_1, a_2, b_2}{\arg\max} \sum_{i=1}^{n} \log \left(q f_1(r_i) + (1-q) f_2(r_i) \right). \tag{16}$$

The estimated parameters for the marginal density f_S of the solar power supply are the solution of an analogous maximization problem.

Note that the maximization problem stated in (16) is difficult to solve because of its high dimensionality. The iterative *expectation-maximization algorithm* (EM algorithm) proposed in Dempster et al. (1977) is particularly appropriate to deal with the maximization of such functions as it enables us to decompose the maximization problem stated in (16) into easier sub-problems. For further details regarding the EM algorithm, see Hastie et al. (2009), Leisch (2004).

3.2.2. Fitting the copula function

To estimate the copula parameter θ, see Table 1, the *inference function for margins method* proposed in Joe and Xu (1996) is applied. The first step of this method is to estimate the marginal distributions, as described in Section 3.2.1. In the second step, the copula parameter θ is fitted using the MLE principle based on the previously estimated marginal distributions.

Indeed, for each copula type in Table 1, the copula density $c^{(\theta)}$ can be obtained by differentiating Eq. (11) in dependence of the copula parameter θ. Using Eq. (8) and the previously fitted marginal distributions we compute the joint density $f_{(R,S)}^{(\theta)}$ for each copula type in dependence of θ. Then, we apply the MLE method to fit the joint density $f_{(R,S)}^{(\theta)}$ for each copula type to the data. As a result, we get an estimate of θ and the corresponding maximum of the product of likelihood functions for each copula type. We choose the copula type and its copula parameter, which gives the largest maximum.

Note that the inference function for margins method represents a certain break with the classical MLE principle, where all model parameters, including the parameters of the marginal distributions, are simultaneously estimated. As an alternative, in Joe (2014) it is proposed to use the components of parameter vector β determined by the two-step procedure in Section 3.2.1 as initial values for iterative numerical methods, which estimate all model parameters simultaneously. But this has not been done in the present paper.

4. Results

To begin with, in Section 4.1, we present the results which we obtained when fitting the model to the pre-processed data at the feed-in points. In Section 4.2, the performance of the prediction model using various validation scores are studied. Section 4.3 quantifies the economic value of the prediction model based on the value score. In Section 4.4, the performance of the copula model are compared to the ones of a quantile regression model. Finally, we check the performance of the fitted models for communities in Section 4.5.

4.1. Fitted model characteristics

The idea of the model fitting procedure was described in Section 3.2. To illustrate the results, which we obtained for the fitted model, we consider an exemplary feed-in point ℓ_0. The marginal distributions of GHI forecasts and the copula parameter are fitted for each hour of the day and forecast lead time separately, whereas the marginal distributions of solar power supply are fitted for each hour of the day. Both datasets are spanning over May, June and July of the years 2015 and 2016, 11-12 UTC. Thus, each of the datasets has about 180 timestamps.

4.1.1. Fitted marginal densities

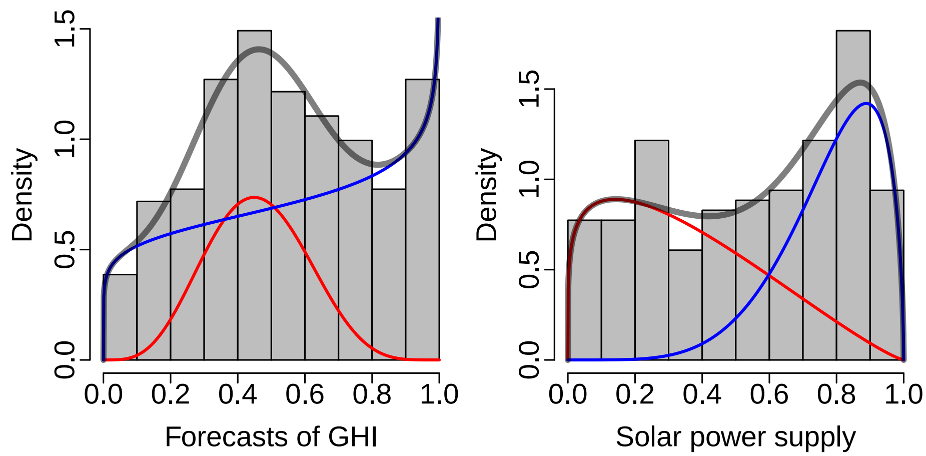


Figure 4: Histograms of rescaled GHI forecasts and rescaled solar power supply with fitted mixed beta density (grey) and weighted component distributions (blue and red) for feed-in point ℓ_0.

The histograms in Figure 4 visualize the empirical distribution of GHI forecasts and solar power supply for the feed-in point ℓ_0. For data of length n the k equally distant bins of both histograms are determined by the *Sturges' rule*, see Scott (2011), i.e.,

$$k = \lceil 1 + \log_2(n) \rceil. \tag{17}$$

If we sum up both suitably weighted curves we get the grey curve in Figure 4, which is the fitted mixed beta density.

From visual inspection of the empirical marginal distributions of the random variables R and S, we observe two modes in the histograms. We see that the fitted marginal densities approximate the shapes of both histograms quite accurately. Therefore, from a qualitative point of view a mixed distribution should be applied.

A quantitative assessment is also provided with the computation of the *Akaike information criterion* (AIC) for five different distribution types (for each feed-in point), see Figure 5. The AIC compares the fit for different distribution types while penalizing choices with a larger number of fitted parameters. It is defined by

$$AIC = 2k - 2\log(L), \tag{18}$$

where k is the number of fitted parameters. The smaller the AIC the better the fits of the distribution. Figure 5 shows that the mixed beta distribution fits the GHI and solar power supply better than the other considered distribution types.

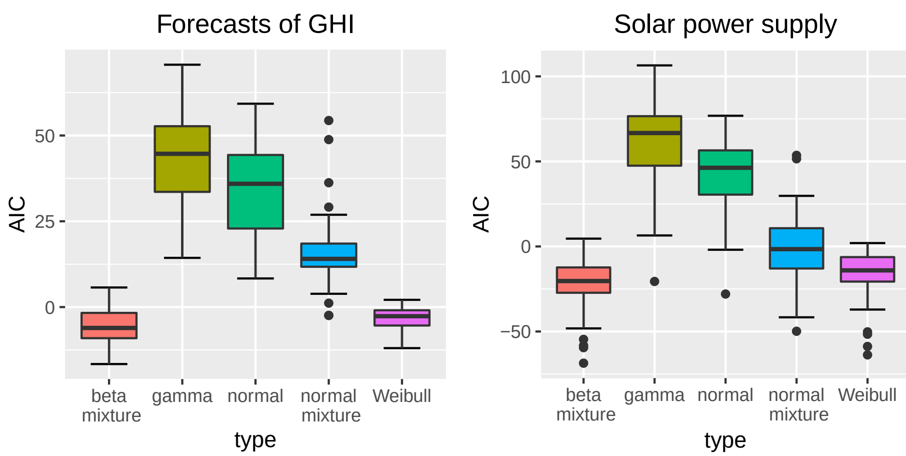


Figure 5: AIC computed for different distribution types fitted to GHI forecasts and solar power supply.

4.1.2. Variation in the marginal densities

Let R_1 and R_2 denote random variables with the beta densities f_1 and f_2, respectively, considered in Eq. (5), likewise for S_1 and S_2. Thus, the random variables R_1 and R_2 are related to the two underlying beta densities of the random variable R, and S_1 and S_2 to the ones of the random variable S. For all feed-in points we computed the expectation and variance of the random variables R_i and S_i using the equations

$$\mathbb{E}(Z) = \frac{a_i}{a_i + b_i}, \tag{19}$$

$$\mathrm{Var}(Z) = \frac{a_i b_i}{(a_i + b_i + 1)(a_i + b_i)^2}, \tag{20}$$

where Z is a random variable designating either R_i or S_i, and $a_i, b_i > 0$ are the shape parameters introduced in Section 3.1.1. The computed expectations and variances over the entire ROI are summarized in Figure 6.

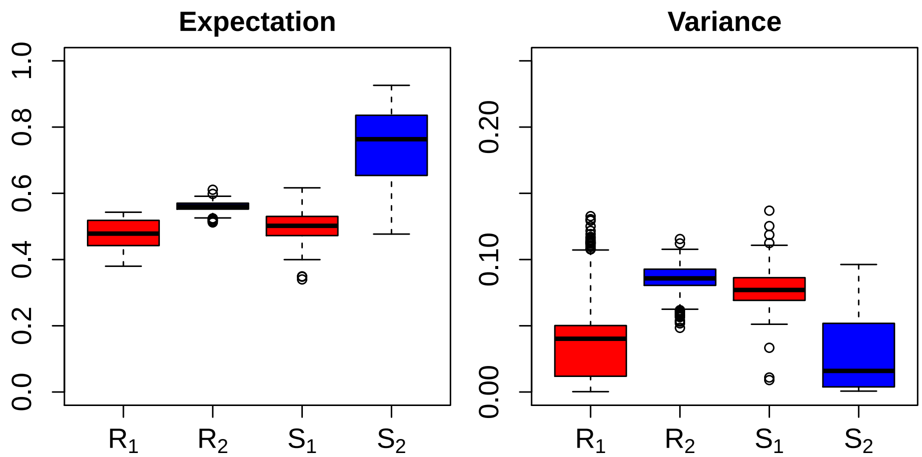


Figure 6: Expectation and variance of the fitted component beta distributions.

Note that the width of most of the boxplots is quite large. This indicates that applying averages of the density parameters over all feed-in points might introduce inaccuracies into the model. Thus, the fitting of the density parameters should be done for each feed-in point separately as it was already mentioned in Section 3.2.1.

4.1.3. Fitted copula and two-dimensional density function

As described in Section 3.2.2, we computed the copula parameter θ and the maximum of the log-likelihood function for each copula type and feed-in point separately. The results are depicted in Table 2 for feed-in point ℓ_0. Table 2 shows that the Frank copula has the largest maximum of the log-likelihood function. Thus, for ℓ_0 the Frank copula performs better than the other copula types considered in Table 1.

copula type	Clayton	Frank	Gumbel	Joe
θ	3.50	7.26	2.94	3.21
$\log\ L(\theta)$	70.63	93.87	69.37	49.08

Table 2: Estimates of the parameter θ and maxima of the log-likelihood function for the exemplary feed-in point ℓ_0.

Altogether, for 166 feed-in points out of the 168 considered feed-in points, see Section 2.1, the Frank copula was determined as best fit while for only two feed-in points a different copula type performed better. Therefore, in order to reduce computational complexity, we decided to apply the Frank copula to all feed-in points.

The copula parameter θ expresses how strongly the random variables R and S are correlated. Note that for each feed-in point ℓ, *Spearman's rank correlation coefficient* $rc(\ell)$ of R and S can be written as

$$rc(\ell) = 12 \int_0^1 \int_0^1 C_\ell(u,v)\mathrm{d}u\mathrm{d}v - 3, \tag{21}$$

where C_ℓ is the copula function fitted to the data at ℓ, see Schweizer and Wolff (1981). In Figure 7 the local Spearman's rank coefficients are visualized. High correlations, but also spatial disparities, can be observed.

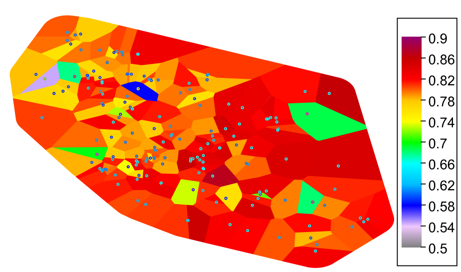


Figure 7: Local Spearman's rank coefficients computed based on fitted copula parameters.

Figure 8 visualizes the fitted joint density $f_{(R,S)}$ and the conditional densities $f_{S|R}$ of solar power supply given a GHI forecast r at the exemplary feed-in point ℓ_0. As expected, with increasing GHI forecasts, the conditional densities assign higher values to larger amounts of solar power supply.

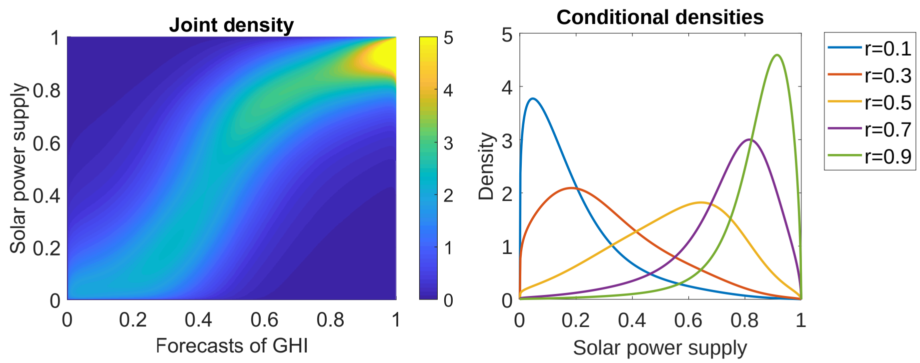


Figure 8: The joint density $f_{R,S}$ and conditional densities $f_{S|R}$ given certain GHI forecasts r for feed-in point ℓ_0.

4.2. Validation based on scores

The proposed copula model was validated using various validation scores, such as the Brier skill score or the continuous ranked probability score, widely applied in the field of weather forecasting, see Wilks (2011). Recall that for model fitting we used datasets spanning over May, June and July of the years 2015 and 2016. For validation we consider the data in the time frame May, June and July of the year 2017.

4.2.1. Notation

In the following we consider GHI forecasts r_i and corresponding measurements of solar power supply s_i where the index i belongs to the validation set $I_{val} = \{1, \ldots, n\}$. We compute the conditional level-crossing probabilities $p_i(v) = P(S \geq v \mid R = r_i)$ for solar power supply exceeding the level $v \in [0, 1]$. The occurrence of the event that solar power supply s_i exceeds the level v is denoted by $o_i(v) = I(v, s_i)$, where I is the indicator function defined as

$$I(v, s) = \begin{cases} 1\,, & \text{if } s \geq v, \\ 0\,, & \text{if } s < v. \end{cases} \tag{22}$$

Thus, for each level v and index $i \in I_{val}$, we consider the pair $(p_i(v), o_i(v))$ of the probability and occurrence of the event $\{S \geq v\}$ for a given feed-in point, hour of the day and forecast time.

4.2.2. Bias

The *bias* of our prediction model is defined as

$$bias(v) = \frac{1}{n}\sum_{i=1}^{n}(p_i(v) - o_i(v)). \tag{23}$$

Note that the quantity $bias(v)$ takes values in $[-1, 1]$ for each $v \in [0, 1]$. In the ideal case, a good prediction model is unbiased, i.e. , $bias(v)$ is equal to zero.

However, an unbiased prediction model does not necessarily generate useful predictions. For instance, if all probabilities are equal to the relative frequency of the occurrences of the considered event, then the predictions are the same for all days. Such a prediction model would be unbiased, but it holds no information in regard to short term changes.

4.2.3. Brier score

Another measure of accuracy is the *Brier score* defined by

$$bs(v) = \frac{1}{n}\sum_{i=1}^{n}(p_i(v) - o_i(v))^2. \tag{24}$$

The Brier score takes its values in the interval $[0, 1]$. It represents the mean squared differences between our predictions and the actual events. Thus, the Brier score of a good prediction model should be near zero.

The Brier score encompasses many important characteristics of a prediction model. Using the algebraic decomposition of the Brier score, more information on the model performance can be obtained (see Wilks (2011)). For a level v, the Brier score $bs(v)$ can be expressed in terms of the reliability rel, resolution res and uncertainty unc, where

$$bs(v) = rel(v) - res(v) + unc(v). \tag{25}$$

The definition of the reliability and resolution requires to partition the unit interval [0,1] into sub-intervals $B_1, \cdots, B_J$. Then, each sub-interval B_j contains

n_j values of forecasts $p_i(v)$ associated to the indicators of the occurring events $o_i(v)$. Furthermore, by $\overline{p}_j(v)$ and $\overline{o}_j(v)$ we denote the mean of the probabilities and the mean of the number of observations for each partition component B_j, i.e.

$$\overline{p}_j(v) = \frac{1}{n_j} \sum_{p_k(v) \in B_j} p_k(v), \tag{26}$$

$$\overline{o}_j(v) = \frac{1}{n_j} \sum_{o_k(v) \in B_j} o_k(v). \tag{27}$$

Moreover, by $\overline{o}(v)$ we denote the *climatological mean* for all observations, i.e.

$$\overline{o}(v) = \frac{1}{n} \sum_{i=1}^{n} o_i(v). \tag{28}$$

Then, the *reliability* is defined as

$$rel(v) = \frac{1}{n} \sum_{j=1}^{J} n_j (\overline{p}_j(v) - \overline{o}_j(v))^2. \tag{29}$$

A small reliability is typical for a well-calibrated prediction model. Besides, the *resolution* is defined as

$$res(v) = \frac{1}{n} \sum_{j=1}^{J} n_j (\overline{o}_j(v) - \overline{o}(v))^2. \tag{30}$$

It measures how large the means of the number of observations for each subinterval differ from the climatological mean for all observations. Higher resolution means that the prediction model is able to distinguish between situations with different frequencies of occurrence. Last but not least, the *uncertainty* is given by

$$unc(v) = \overline{o}(v)(1 - \overline{o}(v)), \tag{31}$$

which summarizes the variability of the observed events. The uncertainty does not depend on the prediction model. A low uncertainty value means that the observed events happen either with high or low frequency.

4.2.4. Brier skill score

The Brier score considered in Section 4.2.3 does not allow for a direct quantitative comparison of the accuracy of prediction models as it depends on the characteristics of the observed event. To draw a clear line between a good and a bad prediction model, the Brier skill score is used. The Brier skill score compares the Brier score $bs(v)$ of the prediction model with the Brier score $bs_r(v)$ of some reference model. Thus, the *Brier skill score* is defined as

$$bss_r(v) = 1 - \frac{bs(v)}{bs_r(v)}. \tag{32}$$

Note that $bss_r(v)$ takes its values in the interval $[-\infty, 1]$. If the reference model gives better results than the actually considered prediction model, the Brier skill score is negative and otherwise positive.

As reference model we consider the climatological model $\overline{o}(v)$ often used as a benchmark for weather forecast models. Since for the climatological model reliability and resolution are equal to zero, the Brier score of the climatological model is equal to the uncertainty *unc* leading to

$$bss_c(v) = \frac{res(v) - rel(v)}{unc(v)}. \tag{33}$$

This implies that $bss_c < 0$ holds if the resolution is smaller than the reliability. This is clearly an undesirable outcome, as the resolution should be high and the reliability small. Therefore, prediction models with $bss_c < 0$ are commonly not considered.

4.2.5. Continuous ranked probability score

We consider the conditional cumulative distribution function of solar power supply $F_{S|R=r_i}$ given a GHI forecast r_i, and the corresponding measured solar power supply s_i. Then, we define the *continuous rank probability score* as

$$crps_i = \int_{-\infty}^{\infty} [F_{S|R=r_i}(x) - F_i(x)]^2 \mathrm{d}x \tag{34}$$

where F_i is the so-called cumulative-probability step function defined as

$$F_i(x) = \begin{cases} 0 & \text{if } x \leq s_i, \\ 1 & \text{if } x > s_i. \end{cases} \tag{35}$$

Furthermore, we compute

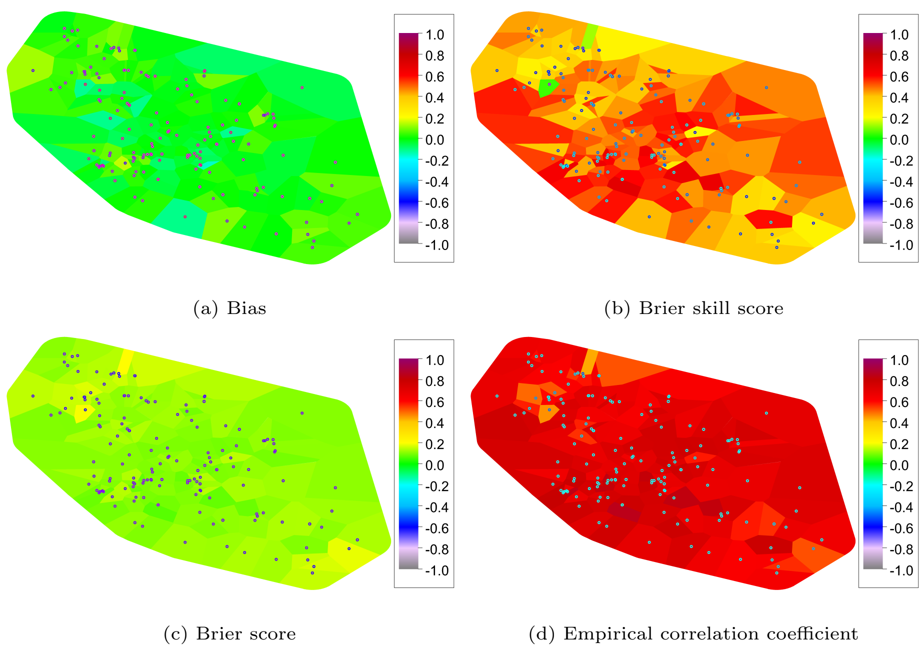
$$crps = \frac{1}{n} \sum_{i=1}^{n} crps_i \tag{36}$$

to quantify how concentrated around the corresponding observations are the computed conditional densities given GHI forecasts. As a general rule we can state: the lower the *crps*, the better.

4.2.6. Validation for each feed-in point

In this section we consider the forecast period 11-12 UTC and the forecast lead time of one hour. We compute the bias, Brier score and Brier skill score for each feed-in point separately. Furthermore, we compute the empirical correlation coefficient $\rho(v)$, see Eq. (2), of the observations $o_i(v)$ and probabilities $p_i(v)$.

The resulting validation scores are visualized in Figure 9 for the threshold $v = 0.8$. The biases and Brier scores are near zero for all feed-in points, whereas almost all computed Brier skill scores and empirical correlation coefficients are high. This indicates that the prediction model proposed in Section 3 works quite well regardless of the considered location.



(a) Bias

(b) Brier skill score

(c) Brier score

(d) Empirical correlation coefficient

Figure 9: Validation scores computed at each feed-in point for the threshold $v = 0.8$

4.2.7. *Validation for different lead times and hours of the day*

It is commonly accepted that the longer the lead time is, the worse is the accuracy of forecasts. This effect is analyzed by merging the validation sets of all feed-in points and applying each validation score considered in Sections 4.2.2 to 4.2.6 to the whole validation set. The analysis becomes then independent of the location and enables us to sum up the information to a single value for each score. Not surprisingly, Table 3 shows that our model yields better results for the shortest lead time of 1h than for longer lead times. However, the validation scores for longer forecast lead times are still highlighting good performance of the proposed model.

lead time (h)	$bias$	bs	bss	ρ	rel	res	unc	$crps$
1	0.006	0.124	0.437	0.661	0.001	0.097	0.221	0.088
4	-0.017	0.152	0.316	0.568	0.006	0.075	0.223	0.105
7	-0.022	0.151	0.315	0.565	0.004	0.073	0.221	0.108
10	-0.035	0.150	0.326	0.578	0.003	0.075	0.222	0.109
13	-0.037	0.141	0.363	0.608	0.003	0.083	0.221	0.106
16	-0.050	0.145	0.346	0.598	0.004	0.080	0.222	0.108
19	-0.069	0.151	0.327	0.590	0.006	0.079	0.224	0.110

Table 3: Validation scores of the combined validation sets for different forecast lead times and the threshold $v = 0.8$.

Finally, we applied our prediction model to GHI forecasts and solar power supply for different hours of the day. Table 4 shows very good validation scores for each 1h-period, regardless of the considered hour of the day. Regarding most scores we get slightly worse results for 13-14 UTC and 17-18 UTC, but even in these cases the brier skill score is clearly greater than zero and the bias is almost zero.

hour of the day	*bias*	*bs*	*bss*	ρ	*rel*	*res*	*unc*	*crps*
5-6	0.005	0.085	0.318	0.565	0.002	0.041	0.124	0.086
6-7	-0.019	0.125	0.415	0.646	0.002	0.090	0.214	0.092
7-8	-0.025	0.137	0.419	0.650	0.002	0.100	0.236	0.094
8-9	0.006	0.146	0.369	0.624	0.006	0.090	0.232	0.093
9-10	-0.002	0.155	0.308	0.570	0.005	0.072	0.224	0.097
10-11	-0.006	0.155	0.291	0.552	0.004	0.067	0.218	0.100
11-12	0.006	0.124	0.437	0.661	0.001	0.097	0.221	0.088
12-13	-0.026	0.147	0.347	0.593	0.002	0.078	0.225	0.098
13-14	-0.028	0.157	0.278	0.535	0.003	0.062	0.217	0.108
14-15	-0.026	0.126	0.397	0.633	0.001	0.084	0.209	0.097
15-16	-0.025	0.127	0.368	0.609	0.002	0.074	0.200	0.100
16-17	0.001	0.097	0.290	0.538	0.002	0.041	0.136	0.097
17-18	-0.060	0.104	0.318	0.601	0.007	0.053	0.153	0.168

Table 4: Validation scores of the combined validation sets for 1h forecast lead time, for different hours of the day (in UTC) and the threshold $v = 0.8$.

4.3. Economic value of the forecast model

Distribution network operators have high interest in the economic value of prediction models. For that purpose, the value score VS is often used to analyze the economic value of a forecast model compared to the climatological model, see Wilks (2011).

The computation of the value score is based on the so-called cost-loss ratio. In our case, the cost C is the cost of a curtailment which should be done if the solar plant generates more solar power supply than the predefined threshold. The loss L corresponds to the averaged economical damage caused by the overloading event when the predefined threshold is exceeded. Then, the *cost-loss ratio* is given by the quotient $C/L \in [0, 1]$. It enables us to characterize various possible scenarios for which the value score can be computed.

To formally define the value score, the relative joint frequencies $p_{0,0}, p_{1,0}, p_{0,1}$

and $p_{1,1}$ given in Table 5 have to be considered.

$p_{1,1} = \dfrac{\#\{i : p_i > C/L, o_i = 1\}}{m}$	$p_{1,0} = \dfrac{\#\{i : p_i > C/L, o_i = 0\}}{m}$
$p_{0,1} = \dfrac{\#\{i : p_i \leq C/L, o_i = 1\}}{m}$	$p_{0,0} = \dfrac{\#\{i : p_i \leq C/L, o_i = 0\}}{m}$

Table 5: Relative joint frequencies considered for the computation of the value score.

Then, the *value score* VS is defined as

$$VS = \begin{cases} \dfrac{(C/L)(p_{1,1} + p_{1,0} - 1) + p_{0,1}}{(C/L)(\overline{o} - 1)}, & \text{if } C/L < \overline{o}, \\ \dfrac{(C/L)(p_{1,1} + p_{1,0}) + p_{0,1} - \overline{o}}{\overline{o}((C/L) - 1)}, & \text{if } C/L > \overline{o}. \end{cases} \tag{37}$$

For more details regarding the interpretation of the value score, see Wilks (2011).

Following the approach of Wilks (2001), we compute the value score for the cost/loss ratio values equal to $(p_i + p_{i+1})/2$, where $i \in \{1, \ldots, m-1\}$. The evolution of the value score with respect to the cost-loss ratio is illustrated in Figure 10. The value score is positive for all considered cost/loss ratios which means that the proposed model is more valuable than the climatological model. Moreover, the proposed model shows a greater economic utility for decision making with value scores larger than 0.5 if $C/L \in [0.12, 0.55]$.

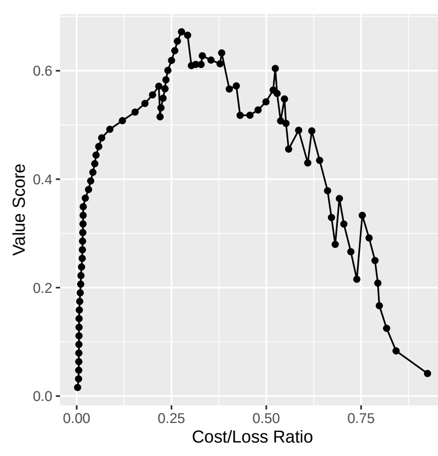


Figure 10: Value score with respect to cost-loss ration for the exemplary feed-in point ℓ_0.

4.4. Comparison with the quantile regression model

The accuracy of the copula model proposed in this paper is finally compared to the one of the quantile regression (QR) model. QR is a non-parametric approach to estimate conditional quantiles of a random variable, called predictand, given some independent random variables, called predictors. Linear QR assumes a linear relationship between the conditional quantiles of the predictand and the predictors, see Davino et al. (2013) for more details.

We consider the time period $11 - 12$ UTC and a forecast lead time of one hour. For each feed-in point, the solar power supply S is the predictand and the global horizontal irradiation R the unique predictor. Then, the conditional α-quantile $\widehat{q}_\alpha(r) = \widehat{b}_\alpha r$ is computed by minimizing the expression

$$\sum_{i=1}^{n} \rho_\alpha(s_i - \widehat{b}_\alpha r_i), \tag{38}$$

where the quantile loss function is defined as

$$\rho_\alpha(u) = \begin{cases} \alpha u, & \text{if } u \geq 0, \\ (\alpha - 1)u, & \text{otherwise.} \end{cases} \tag{39}$$

By applying 101 linear quantile regressions, i.e. with $\alpha \in \{0, 1/100, \ldots, 1\}$, we compute conditional α-quantiles for solar power supply given GHI forecasts. Next, for a threshold v we approximate the conditional level-crossing probability of S given the GHI forecast r under the quantile regression model by

$$\pi(v) = 1 - \min_\alpha(\widehat{q}_\alpha(r) - v). \tag{40}$$

The reliability diagrams are then computed for the copula model and the quantile regression. Note that the corresponding reliability diagrams are the pairs of points $(\overline{p}_j(v), \overline{o}_j(v))$ and $(\overline{\pi}_j(v), \overline{o}_j(v))$ for $j \in \{1, \ldots, J\}$, where $\overline{p}_j(v)$ and $\overline{\pi}_j(v)$ are the mean conditional level-crossing probabilities of S given R under the copula model and the QR model, respectively, for the sub-interval B_j (using the same notation as in Section 4.2.3).

For a perfectly reliable model, the difference between the mean conditional probability and the corresponding frequency of occurrence of the considered event is zero, i.e., the reliability diagram coincides with the diagonal (grey line in Figure 11). Vice versa, the more the reliability diagram deviates from the diagonal, the less reliable the model, see Wilks (2011) for further details.

Figure 11 compares the reliability diagrams of the two models. The numbers of samples n_j in the sub-intervals B_j are visualized in bars below the diagrams. We observe that the copula model is more reliable than the quantile regression model which tends to underestimate the probability that the solar power supply exceeds the predefined threshold $v = 0.8$.

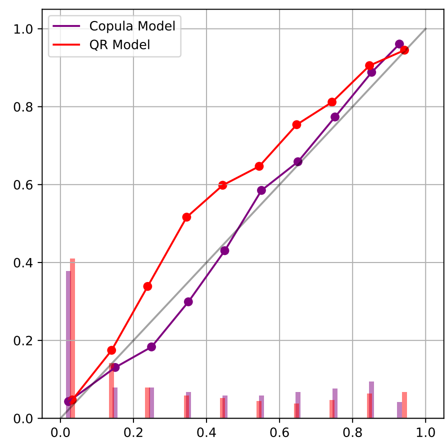
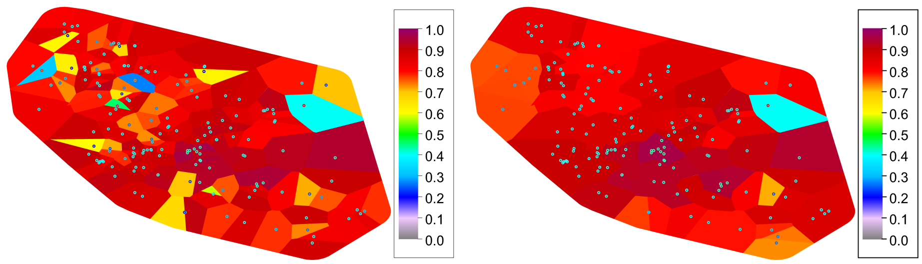


Figure 11: Reliability diagrams of the copula model (in violet) and the QR model (in red), for the predefined threshold $v = 0.8$.

4.5. Comparison between different hierarchy levels

Since overloading problems can occur at each hierarchy level of a distribution network, it is crucial that probabilistic predictions quantifying the risk of overloading can be generated for several hierarchy levels. In fact, the consequences of a critical event are usually even larger the higher the hierarchy level is, where it occurs. Indeed, in such a case the solar power supply of a whole region might be interrupted. Consequently, it is convenient to have a flexible prediction model, which can be applied to different hierarchy levels of a distribution network.

Using the methods stated in Section 3.1 we compute and visualize conditional level-crossing probabilities for a certain threshold at feed-in points and communities. It turns out that there are less local disparities for communities than for feed-in points. In fact the aggregation of solar power supply causes a smoothing effect on the conditional probabilities, see Figure 12.



(a) Probabilities for feed-in points

(b) Probabilities for communities

Figure 12: Conditional level-crossing probabilities for solar power supply exceeding the threshold of 0.8 for May 10, 2017, 11-12 UTC.

Table 6 clearly shows the flexibility of our prediction model by comparing the averaged validation scores computed for communities with the averaged scores we get for individual feed-in points. Some scores for communities are even slightly better than for feed-in points. This stands to reason, since averaging over a certain number of feed-in points eliminates noise.

hierarchy level	*bias*	*bs*	*bss*	ρ	*rel*	*res*	*unc*	*crps*
feed-in points	0.006	0.124	0.437	0.661	0.001	0.097	0.221	0.088
communities	0.004	0.113	0.491	0.701	0.002	0.110	0.223	0.080

Table 6: Validation scores of the combined validation sets for feed-in points and communities with lead time one hour and forecast period 11-12 UTC.

5. Conclusion

In this paper a probabilistic prediction model to quantify the risk of overloading at feed-in points was proposed. The model is based on hourly deterministic GHI forecasts and applies copula theory to compute the joint distribution of GHI forecast and solar power supply for each feed-in point. Based on marginal and joint distributions, conditional probabilities for solar power supply exceeding a predefined threshold are computed.

The model was validated using prediction scores, such as bias, Brier score (decomposed into reliability, resolution and uncertainty), Brier skill score, continuous rank probability score and the empirical correlation coefficient. The scores were validated for forecast lead times ranging from 1 to 19 hours and 1h periods of the day ranging from 5 to 18 UTC. These validations showed a high accuracy of the proposed copula-based model, regardless of the considered hour of the day or forecast lead time. Moreover, a comparison of the copula model reliability diagram with the one of the quantile regression model emphasized higher reliability than a state-of-the-art model. Besides, we also showed that the model can be applied to higher hierarchy levels in the distribution network, such as communities. Finally, the value score was computed to quantify the economic value of the proposed model for an exemplary feed-in point. The copula model demonstrated a higher economic utility than the climatological model for the selected feed-in point, regardless of the cost-loss ratio values. Particularly, the economic utility was outstanding for cost-loss ratios in the interval $[0.12, 0.55]$.

The fitting and the validation of the copula model have been undertaken on three typical months with high GHI in Germany (May, June and July). If more than three months have to be considered, it may be necessary to split the fitting and validation period in order to capture information relevant to seasonal variation. The way to split the data, if necessary, is not straightforward and may have to be undertaken by trial-and-error.

Besides, it may also be possible to integrate other factors, except GHI, that are correlated to the solar power supply. This will be investigated in a forthcoming paper using more advanced tools of copula theory such as nesting or vine copulas (Joe, 2014).

Acknowledgments

We would like to thank "Bundesministerium für Bildung und Forschung" (BMBF) for financially supporting this research project (grant 05M18VUB).

References

Alessandrini, S., Delle Monache, L., Sperati, S., Cervone, G., 2015. An analog ensemble for short-term probabilistic solar power forecast. Applied Energy 157, 95–110.

Almeida, M.P., Perpinan, O., Narvarte, L., 2015. PV power forecast using a nonparametric PV model. Solar Energy 115, 354–368.

Antonanzas, J., Osorio, N., Escobar, R., Urraca, R., Martinez-de Pison, F., Antonanzas-Torres, F., 2016. Review of photovoltaic power forecasting. Solar Energy 136, 78–111.

Bacher, P., Madsen, H., Nielsen, H.A., 2009. Online short-term solar power forecasting. Solar Energy 83, 1772–1783.

Baldauf, M., Seifert, A., Förstner, J., Majewski, D., Raschendorfer, M., Reinhardt, T., 2011. Operational convective-scale numerical weather prediction with the COSMO model: Description and sensitivities. Monthly Weather Review 139, 3887–3905.

Bessa, R.J., Trindade, A., Silva, C.S., Miranda, V., 2015. Probabilistic solar power forecasting in smart grids using distributed information. International Journal of Electrical Power & Energy Systems 72, 16–23.

Bird, L., Cochran, J., Wang, X., 2014. Wind and solar energy curtailment: Experience and practices in the United States. Technical Report. National Renewable Energy Laboratory.

Coiffier, J., 2011. Fundamentals of Numerical Weather Prediction. Cambridge University Press.

Council of European Energy Regulators, 2017. Ceer report on power losses. URL: `https://www.ceer.eu/documents/104400/-/-/09ecee88-e877-3305-6767-e75404637087`.

Davino, C., Furno, M., Vistocco, D., 2013. Quantile Regression: Theory and Applications. volume 988. J. Wiley & Sons.

Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological) 39, 1–22.

Dickert, J., Hable, M., Schegner, P., 2009. Energy loss estimation in distribution networks for planning purposes, in: IEEE Bucharest PowerTech, pp. 1–6.

Durante, F., Sempi, C., 2015. Principles of Copula Theory. Chapman and Hall/CRC.

Golestaneh, F., Gooi, H.B., 2017. Multivariate prediction intervals for photovoltaic power generation, in: 2017 IEEE Innovative Smart Grid Technologies-Asia (ISGT-Asia), IEEE. pp. 1–5.

Golestaneh, F., Gooi, H.B., Pinson, P., 2016a. Generation and evaluation of space–time trajectories of photovoltaic power. Applied Energy 176, 80–91.

Golestaneh, F., Pinson, P., Gooi, H.B., 2016b. Very short-term nonparametric probabilistic forecasting of renewable energy generation—with application to solar energy. IEEE Transactions on Power Systems 31, 3850–3863.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer.

Heinemann, D., Lorenz, E., Girodo, M., 2006. Solar Energy Resource Management for Electricity Generation from Local Level to Global Scale. Nova Science Publishers.

Hess, R., Glashoff, J., Reichert, B., 2015. The Ensemble-MOS of Deutscher Wetterdienst, in: EMS Annual Meeting Abstracts, Sofia.

Huang, J., Perry, M., 2016. A semi-empirical approach using gradient boosting and k-nearest neighbors regression for gefcom2014 probabilistic solar power forecasting. International Journal of Forecasting 32, 1081–1086.

Hämmerlin, G., Hoffmann, K., 2012. Numerical Mathematics. Springer.

Jobson, J., 1991. Applied Multivariate Data Analysis: Regression and Experimental Design. Springer.

Joe, H., 2014. Dependence Modeling with Copulas. Chapman and Hall/CRC.

Joe, H., Xu, J., 1996. The estimation method of inference functions for margins for multivariate models. URL: https://open.library.ubc.ca/collections/facultyresearchandpublications/52383/items/1.0225985.

Kaldellis, J., Kapsali, M., Kavadias, K., 2014. Temperature and wind speed impact on the efficiency of PV installations. Experience obtained from outdoor measurements in greece. Renewable Energy 66, 612–624.

Karimi, M., Mokhlis, H., Naidu, K., Uddin, S., Bakar, A., 2016. Photovoltaic penetration issues and impacts in distribution network - A review. Renewable and Sustainable Energy Reviews 53, 594–605.

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

Leisch, F., 2004. A general framework for finite mixture models and latent glass regression in R. Journal of Statistical Software, 11 (8) , 1–18.

Lu, Q., Hu, W., Min, Y., Yuan, F., Gao, Z., 2014. Wind power uncertainty modeling considering spatial dependence based on pair-copula theory, in: PES General Meeting— Conference & Exposition, IEEE. pp. 1–5.

Massidda, L., Marrocu, M., 2018. Quantile regression post-processing of weather forecast for short-term solar power probabilistic forecasting. Energies 11, 1763.

Mekhilef, S., Saidur, R., Kamalisarvestani, M., 2012. Effect of dust, humidity and air velocity on efficiency of photovoltaic cells. Renewable and Sustainable Energy Reviews 16, 2920–2925.

Nelsen, R., 2006. An Introduction to Copulas. Springer.

Panamtash, H., Zhou, Q., Hong, T., Qu, Z., Davis, K.O., 2020. A copula-based Bayesian method for probabilistic solar power forecasting. Solar Energy 196, 336–345.

Papaefthymiou, G., Kurowicka, D., 2009. Using copulas for modeling stochastic dependence in power system uncertainty analysis. IEEE Transactions on Power Systems 24, 40 – 49.

Schweizer, B., Wolff, E., 1981. On nonparametric measures of dependence for random variables. The Annals of Statistics 9, 879–885.

Scott, D.W., 2011. Sturges' and Scott's Rules. Springer. pp. 1563–1566. URL: https://doi.org/10.1007/978-3-642-04898-2_578, doi:10.1007/978-3-642-04898-2_578.

SolarPower Europe, 2017. Global market outlook 2018-2022. URL: http://www.solarpowereurope.org/wp-content/uploads/2018/09/Global-Market-Outlook-2018-2022.pdf.

Vale, P., 2015. Energy assessment of photovoltaic conversion systems. Technical Report. Instituto Superior Técnico.

Wang, Y., Infield, D., Stephen, B., Galloway, S., 2014. Copula-based model for wind turbine power curve outlier rejection. Wind Energy 17, 1677–1688.

Wilks, D., 2001. A skill score based on economic value for probability forecasts. Meteorological Applications 8, 209–219.

Wilks, D.S., 2011. Statistical Methods in the Atmospheric Sciences. Academic Press.

Zamo, M., Mestre, O., Arbogast, P., Pannekoucke, O., 2014. A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production. Part II: probabilistic forecast of daily production. Solar Energy 105, 804–816.

Zhang, B., Dehghanian, P., Kezunovic, M., 2016. Spatial-temporal solar power forecast through use of gaussian conditional random fields, in: IEEE Power and Energy Society General Meeting (PESGM), pp. 1–5.