Noname manuscript No.
(will be inserted by the editor)

Probabilistic analysis of solar power supply using D-vine copulas based on meteorological variables

Freimut von Loeper · Tom Kirstein · Basem Idlbi · Holger Ruf · Gerd Heilscher · Volker Schmidt

Received: date / Accepted: date

Abstract Solar power generation at solar plants is a strongly fluctuating non-deterministic variable depending on many influencing factors. In general, it is not clear which and how certain variables influence solar power supply at feed-in points in a distribution network. Therefore, analyzing the dependence structure of measured solar power supply and other variables is very informative and can be helpful in designing probabilistic prediction models.
In this paper multivariate D-vine copulas are fitted to investigate the relationship between solar power supply and certain meteorological variables in the current time period of one hour length as well as solar power supply in previous time periods. The meteorological variables considered in this analysis are global horizontal irradiation, temperature, wind speed, humidity, precipitation and pressure. By applying parametric D-vine copulas useful insight is gained into the dependence structure of solar power supply and the considered meteorological variables. The main goal lies in determining suitable explanatory variables for the design of probabilistic prediction models for solar power supply at single feed-in points and analyzing their impact on the validation of conditional level-crossing probabilities.

Keywords D-vine copula · dependence structure · solar power supply · meteorological variable · conditional level-crossing probability

1 Introduction

In recent years global warming was acknowledged as a serious problem becoming a topic of public concern. To reduce carbon dioxide emissions and limit

Freimut von Loeper
Helmholtzstraße 18, 89069 Ulm
Tel.: +49 731/50-23528
Fax: +49 731/50-23649
E-mail: freimut.von-loeper@uni-ulm.de

further global warming, alternative energy sources such as renewable energy are required. For that reason, the renewable energy sector has the support of the governments in many countries and is growing rapidly [4]. Especially the solar energy sector has been reporting record growth for many years [24].

However, higher solar power penetration might lead to new problems for distribution network operators. Since solar power generation strongly depends on weather conditions, the prospective solar power supply of the distribution network cannot be easily taken into account. Thus, it might be difficult to predict excessive power flow in the grid and to prevent voltage violations and overloading of power lines and transformers which destabilize the electricity supply and cause economic damages [14]. To regulate fluctuations in power and load of distribution networks, automated and more economic applications for smart technologies are needed [5]. Smart grid management can use solar power forecasting with hourly forecasting horizons for power system operation such as economic dispatch and unit commitment [26].

To compute probabilistic predictions regarding the generation of solar power supply at certain feed-in points, several high-dimensional stochastic models are considered in the literature. In [2] a non-parametric quantile regression forest is used to predict solar power supply based on several meteorological variables such as temperature, wind speed, humidity, sea level pressure and cloud cover at different levels. The prediction model proposed in [12] takes solar radiation, temperature, cloud ice water content and wind speed as input parameters to compute prediction intervals based on k-nearest neighbor regression. In [3] wind and solar power are computed by applying a combination of the gradient boosting tree algorithm and feature engineering techniques, where the authors of [3] concluded that information about the forecast grid further improves the prediction results. In [28] Gaussian conditional random fields are used to model the spatial-temporal dependence at neighboring feed-in points. However, the papers mentioned above give little insight into the dependence structure of solar power supply and meteorological variables. In particular, it is not clear which explanatory variables are suitable for probabilistic prediction of solar power supply.

An alternative approach can be given by multivariate copulas which are applied to compute the joint distribution of interdependent random variables [19]. In the literature of renewable energy modeling copulas are mostly used to account for spatial, temporal or spatio-temporal dependence at neighboring wind parks or feed-in points of solar power. In [18] and [20] Gaussian copulas and R-vine copulas are applied to model the spatial dependence of the wind power supply generated at many different wind parks. In [6] and [10] Gaussian and D-vine copulas are utilized to model the temporal and spatio-temporal dependence of these characteristics. Furthermore, in [8] and [9] Gaussian and R-vine copulas are applied to model the spatial dependence of solar power supply at neighboring feed-in points. However, multivariate copulas are not only helpful to analyze and model the spatial or temporal dependence of wind or solar power supply at neighboring feed-in points, but also the relationship

between power supply and other influencing factors, such as meteorological variables.

In the present paper, we determine suitable explanatory variables for the design of probabilistic prediction models for solar power supply at single feed-in points, extending the methodology which has recently been used in [17]. For that purpose, multivariate D-vine copulas are applied to analyze the dependence structure of the considered meteorological variables and the solar power supply in the current time period of one hour length as well as in previous hourly periods. The stepwise fitting of D-vine copulas helps us in analyzing the interdependence of the considered variables in detail. Moreover, conditional level-crossing probabilities are computed and validated with prediction scores to determine suitable explanatory variables for the design of prediction models which compute probabilities for critical amounts of solar power supply. Last but not least the effect of the considered month and time of day on the probabilistic prediction of solar power supply is investigated.

The rest of the paper is organized as follows. In Section 2 the data is described and analyzed empirically. Section 3 introduces the mathematical methodology applied in this paper. In Section 4 the results are presented and discussed. Section 5 concludes.

2 Data

The measurements of solar power supply and reverse power utilized in this paper were collected in cooperation with the local distribution network operator Stadtwerke Ulm/Neu-Ulm Netze GmbH (SWUN). Additional meteorological information concerning global horizontal irradiation (GHI) was gathered by satellites where temperature, wind speed, humidity, precipitation and pressure were computed by reanalysis and published as an open source, see [7, 23] respectively. The datasets of both sources are described in Section 2.1 and empirically analyzed in Section 2.2.

2.1 Data description

Solar power supply of a community near the city of Ulm, called Hittistetten, is measured for the years 2016-2018. The location is a test site defined on the website of SWUN [25]. The power measurements are taken by smart meters at 14 PV systems in Hittistetten. Afterwards they are summed over all PV systems in the community to get the representative solar power supply of the community which is considered in the following. Furthermore, power measurements at the local area transformers are evaluated and daily accumulated reverse energy is estimated. Note that there exist some gaps in the collected data, because of some possible errors in the smart meter infrastructure.

The hourly measurements of solar power supply, the hourly meteorological data and the daily accumulated reverse energy measurements are rescaled to

variable	physical meaning
P_t	solar power supply in the period t hours before the current time period of one hour length
M_1	GHI at ground level in the current period
M_2	temperature at 2 meters above ground in the current period
M_3	precipitation at ground level in the current period
M_4	relative humidity at 2 meters above ground in the current period
M_5	wind speed at 10 meters above ground in the current period
M_6	pressure at ground level in the current period

Table 1 Considered variables and their physical meaning.

the unit interval using a linear transformation. All time stamps are converted to Central European Time (CET). The min-max-rescaled measurements of hourly aggregated solar power supply and data concerning meteorological variables are interpreted as realizations of random variables, see [17] for details. In Table 1, the mathematical symbols P_t and M_i are introduced for the random variables considered in the present paper. Furthermore, rescaled power exceeding 70 percent of the highest recorded amount of solar power supply is seen as a critical amount of solar power supply. Therefore, in the following the value of 0.7 is considered as an exemplary high threshold of rescaled power supply.

2.2 Empirical data analysis

In Figure 1, the daily accumulated reverse energy is visualized. The box plots indicate that during the winter months of January, February, November and December the amounts of reverse power are very low in comparison to the remaining months. In particular, there are no critical overloading events for the distribution network happening within these months. Therefore, since these months are not interesting for the distribution network operators, they are excluded in the present study. Furthermore, in Figure 2 the means of hourly aggregated solar power supply are plotted for each time of day. For all months considered in Figure 2 the means of the hourly aggregated solar power supply are always below the threshold of 0.7 for the time periods of 6-9 CET and 16-18 CET. Hence, our analysis is limited to the months from March to August and the time of day from 9-16 CET.

For the solar power supply P_0 in the current time period and the meteorological variables $M_1, \ldots, M_6$, the panels in Figure 3 show pairwise scatter plots, Kendall's rank correlation coefficients and histograms of these variables. In particular, in the panels below the diagonal the pairwise scatter plots provide detailed information about the kind of pairwise dependence between the

considered variables, which is further quantified by pairwise Kendall's rank correlation coefficients in the panels above the diagonal. Note that the Kendall's rank correlation coefficient measures the rank adjusted correlation between two random variables X and Y and can be estimated by

$$\widehat{\tau} = \frac{2}{n(n-1)} \sum_{i<j} \operatorname{sgn}(x_i - x_j) \operatorname{sgn}(y_i - y_j) \tag{1}$$

for given realizations $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ of X and Y, respectively.

Figure 3 shows that the meteorological variables temperature (M_2), precipitation (M_3), and humidity (M_4) are stronger correlated with solar power supply than wind speed (M_5) and pressure (M_6). Furthermore, the histograms in the panels on the diagonal give a rough idea about the shape of the density of the corresponding variables. Solar power supply (P_0), GHI (M_1) and temperature (M_2) might have a bimodal distribution, whereas the distributions of precipitation (M_3), humidity (M_4), wind speed (M_5) and pressure (M_6) appear to be unimodal.

In Figure 4, the Kendall's rank correlation coefficients of the measurements of solar power supply in the current period of one hour length and in certain previous hourly periods are plotted. As expected the correlation gets weaker the larger the time difference is between both measurements.

In the following analysis our focus is put on variables which have a rank correlation coefficient larger than 0.25 to solar power supply in the current period. The remaining variables are too weakly correlated to have sufficiently much influence on solar power supply. Based on the correlation plots of Figures 3 and 4 only the variables M_1,M_2,M_3,M_4, P_1, P_2, and P_3 fulfill the requirement mentioned above.

3 Methodology

If several explanatory variables are considered, high pairwise correlation does not necessarily mean that all of them are good explanatory variables. In the worst case it might happen that all helpful information within an explanatory variable is part of other explanatory variables. To analyze the dependence structure of interdependent random variables in-depth, our modeling approach is based on D-vine copulas which will be explained in this section.

3.1 Modeling approach

The main goal of the present paper is to determine those components of the random vector $E = (M_1, M_2, M_3, M_4, P_1, P_2, P_3)$ which are useful explanatory variables for the design of probabilistic prediction models for solar power supply. In particular, their impact on the probabilistic prediction of critical amounts of solar power supply is investigated. That impact is quantified by computing conditional level-crossing probabilities of solar power supply P_0

given that $E = e$ for some realization $e \in \mathbb{R}^7$ of the random vector E. In particular, for the exemplary threshold 0.7 we get that

$$P(P_0 \geq 0.7 \mid E = e) = \int_{0.7}^{1} f_{P_0|E=e}(x)\,dx = \int_{0.7}^{1} \frac{f_{P_0,E}(x,e)}{f_E(e)}\,dx$$
$$= \int_{0.7}^{1} \frac{f_{P_0,E}(x,e)}{\int_0^1 f_{P_0,E}(y,e)\,dy}\,dx = \frac{\int_{0.7}^{1} f_{P_0,E}(x,e)\,dx}{\int_0^1 f_{P_0,E}(y,e)\,dy}.$$

To determine the joint probability density $f_{P_0,E}$ of the random vector (P_0, E) the marginal densities of $P_0, M_1, M_2, M_3, M_4, P_1, P_2$ and P_3 are estimated in a first step and D-vine copulas are applied in a second step. Note that our approach based on D-vine copulas gives us in-depth insight into the dependence structure of solar power supply and the considered meteorological variables.

3.2 D-vine copulas

A function $C : [0,1]^d \to [0,1]$ with $d \geq 2$ is called a *d-dimensional copula* if C is the joint cumulative distribution function (CDF) of a d-dimensional random vector with uniformly distributed marginals in $[0,1]$. Copulas are a powerful tool to parametrically model multivariate CDFs with non-Gaussian marginals. The reason for this is the following fundamental *theorem of Sklar*, see [13,19].

Let $F : [0,1]^d \to [0,1]$ be an arbitrary multivariate CDF of a random vector $(X_1, ..., X_d)$. Then, for the marginal CDFs $F_i(x) = P(X_i \leq x)$ with $i \in \{1, .., d\}$ there exists a d-dimensional copula C such that F can be expressed as

$$F(x_1, .., x_d) = C(F_1(x_1), ..., F_d(x_d)) \qquad \text{for each } (x_1, \ldots, x_d) \in \mathbb{R}^d. \tag{2}$$

Thus, the CDF F of of an arbitrary random vector $X = (X_1, \ldots, X_d)$ can be represented as superposition of the corresponding marginal CDFs $F_1, \ldots, F_d$ and a certain copula C which models the dependence structure of the components $X_1, \ldots, X_d$. The parameters of marginal distributions and copula can be estimated separately from data which reduces the number of parameters fitted simultaneously, see [13,19] for further details.

In the present paper a special class of d-dimensional copulas, so-called D-vine copulas, is used which are a popular type of multivariate copulas, see e.g. [6,10]. We assume that the CDF F has a d-dimensional density $f_{1,\ldots,d}$ and denote the 1-dimensional densities of the corresponding marginal CDFs $F_1, \ldots, F_d$ by $f_1, \ldots, f_d$, respectively. To obtain the D-vine structure the d-dimensional density $f_{1,\ldots,d}$ is decomposed into conditional densities by

$$f_{1:d} = f_{d|1:d-1} \cdots f_{2|1:1} f_1, \tag{3}$$

where $f_{i|1:i-1}$ is the conditional density given by $f_{i|1:i-1} = f_{1,\ldots,i}/f_{1,\ldots,i-1}$ for $i = 2, \ldots, d$. In the next step, Sklar's theorem is applied to the conditional density $f_{i,j|i+1:j-1}$ of the random vector (X_i, X_j) with $i + 1 < j$

given that $X_{i+1:j-1} = x_{i+1:j-1}$, where $X_{i+1:j-1} = (X_{i+1}, \ldots, X_{j-1})$ and $x_{i+1:j-1} = (x_{i+1}, \ldots, x_{j-1})$. Considering probability densities instead of distribution functions on both sides of Equation 2 we get that

$$f_{i,j|i+1:j-1} = c_{i,j|i+1:j-1}(F_{i|i+1:j-1}, F_{j|i+1:j-1}) f_{i|i+1:j-1} f_{j|i+1:j-1}, \quad (4)$$

where $c_{i,j|i+1:j-1}$ is some bivariate copula density, $F_{i|i+1:j-1}$ and $F_{j|i+1:j-1}$ denote the conditional CDFs of X_i and X_j, respectively, given that $X_{i+1:j-1} = x_{i+1:j-1}$, and $f_{i|i+1:j-1}$ and $f_{j|i+1:j-1}$ are the corresponding conditional marginal densities. This results in

$$f_{j|i:j-1} = \frac{f_{i,j|i+1:j-1}}{f_{i|i+1:j-1}} = c_{i,j|i+1:j-1}(F_{i|i+1:j-1}, F_{j|i+1:j-1}) f_{j|i+1:j-1}. \quad (5)$$

Note that for $j = i + 1$ Sklar's theorem is used for the (unconditional) bivariate density of the random vector (X_i, X_j). Finally, Equation 5 is repeatedly applied to the conditional densities on the right-hand side of Equation 3 which leads to

$$f_{1:d} = \prod_{j=1}^{d-1} \prod_{i=1}^{d-j} c_{i,j|i+1:j-1}(F_{i|i+1:j-1}, F_{j|i+1:j-1}) \prod_{k=1}^{d} f_k. \quad (6)$$

To estimate the D-vine copula the copula densities $c_{i,j|i+1:j-1}$ in Equation 6 we assume that they do not depend to be specific values of $x_{i+1:j-1}$. Further mathematical details regarding copula theory and, in particular, D-vine copulas can be found e.g. in [13].

3.3 Fitting procedure

To select parametric families of (marginal) distributions for the components $X_1, \ldots, X_d$ of the random vector $X = (X_1, \ldots, X_d)$ considered in Section 3.2, histograms are computed for visual inspection based on all available data $(x_1, \ldots, x_d)$. Thus, the dimension d and the random variables $X_1, \ldots, X_d$ have to be specified, see Table 2. Their histograms give us a rough idea of the shapes of the densities to be fitted and, in particular, exhibit distributional properties such as multimodality. This information is helpful in determining distribution types which might give a good fit to the data considered in this paper. Once candidates for the distribution types are selected, we compute the *Bayesian information criterion* (BIC) defined as

$$BIC = k \ln d - 2 \ln L \quad (7)$$

for each distribution type with k parameters, where L is the maximized value of the likelihood function. The distribution type with the smallest BIC value is a reasonable choice, see [15]. In the next step the parameters of unimodal distributions are estimated by maximizing the likelihood and the parameters of bimodal distributions are fitted by applying the expectation-maximization algorithm, see [11,16].

Finally, to fit a multivariate D-vine copula two further steps have to be carried out:

1) Select parametric families of copula densities for the bivariate copula densities $c_{i,j|i+1:j-1}$ considered in Equation 6.

2) Estimate the parameter(s) for each bivariate copula density $c_{i,j|i+1:j-1}$ in the D-vine structure.

According to the fitting procedure described above, we first specify the input variables $X_1, \ldots, X_d$, before selecting the types of marginal distributions and fitting their parameters. We put $d = 8$ and consider P_0 and M_1, see Table 1, which are the most important input variables. Secondly, the meteorological input variables M_2, M_3 and M_4 are taken into account. To complete the specification of input variables, the solar power inputs P_1, P_2 and P_3 at previous periods of time are added to the other variables according to their temporal ordering. The chosen specification of all considered variables is summarized in Table 2.

general notation	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
specification	P_0	M_1	M_2	M_3	M_4	P_1	P_2	P_3

Table 2 Specification of input variables in the D-vine copula structure

Finally, the conditional bivariate copula densities considered in Equation 6 are determined sequentially, see [13] for details. In particular, for each bivariate copula the copula type and its parameter(s) are determined by maximizing the corresponding likelihood. As candidates the Archimedean copula types Joe, Frank, Clayton and Gumbel are considered, which guarantee a large variety of possible tail dependencies. A detailed description of the fitting procedure of bivariate copulas can be found e.g. in [17]. The results are obtained using the VineCopula package in R, see [21].

4 Results and discussion

In this section we apply the D-vine copula model, which has been described in Section 3, to a sample of 8-dimensional data vectors, in order to get deeper insight into the dependence structure of solar power supply and the considered meteorological variables.

4.1 Model fitting and validation

The parametric families of beta, mixed beta, log-normal, Weibull and gamma distributions are chosen as candidates for the marginal distributions of the random vector $(P_0, M_1, M_2, M_3, M_4, P_1, P_2, P_3)$, based on the histograms visualized in Figure 3. To determine the most suitable distribution type, the BIC

score given in Equation 7 is computed for each of the 7 considered hours of day separately. In Figure 5, the obtained BIC scores are visualized by means of box-plots. Most of the BIC scores of the mixed beta distribution are clearly smaller than those of other distribution types, in particular, for solar power supply, GHI and precipitation. However, for humidity and temperature most distribution types have similar BIC scores. For all input variables the distribution type with the smallest average BIC is chosen, where averaging is taken over the 7 considered hours of day. As a result, Weibull distributions are fitted for humidity and mixed beta distributions for solar power supply, GHI, precipitation and temperature.

In Figure 6 the fitted marginal densities and the underlying histograms are visualized for solar power supply, GHI, humidity, precipitation and temperature regarding the exemplary hour of day from 14-15 CET. Moreover, to compare the quality of the fits for each input variable the parametric density with the second best average BIC is visualized as well.

Figure 7 shows the D-vine copula structure fitted to the input variables specified in Table 2. Each panel entitled with $i, j|i+1, \ldots, j-1$ visualizes the fitted conditional copula density $c_{i,j|i+1:j-1}$, see Section 3.2. The bivariate copulas with symmetric tail-dependence, see Figure 7, are modelled using the Frank copula, see e.g. the copula entitled with 4,5, whereas the bivariate copulas with assymmetric tail-dependence are modelled using Joe, Clayton or Gumbel copulas, see e.g. the copula entitled with 1,2.

P_0	0.54	0.25	-0.38	-0.38	0.50	0.43	0.43
0.57	M_1	0.34	-0.39	-0.38	0.54	0.47	0.43
0.29	0.36	M_2	-0.11	-0.09	0.26	0.28	0.23
-0.42	-0.38	-0.10	M_3	0.67	-0.46	-0.48	-0.48
-0.41	-0.35	-0.07	0.63	M_4	-0.47	-0.47	-0.44
0.59	0.54	0.31	-0.43	-0.44	P_1	0.64	0.51
0.52	0.48	0.33	-0.47	-0.45	0.64	P_2	0.65
0.49	0.45	0.30	-0.48	-0.44	0.54	0.67	P_3

Table 3 Pairwise Kendall's rank correlation coefficients of original data (below the diagonal) and of realizations drawn from the fitted D-vine copula model (above the diagonal) computed for the input variables on the diagonal of the corresponding rows/columns.

In Table 3 pairwise Kendall's rank correlation coefficients are presented which have been computed based on original data and 5000 simulated realizations drawn from the D-vine copula for the input variables on the diagonal of the corresponding rows/columns, using the algorithm explained in [13, 1] and its implementation in R, see the VineCopula package in [21]. By comparing the Kendall's rank correlation coefficients it becomes apparent that the coefficients of the original data are very similar to the coefficients based on the simulations of the fitted D-vine copula model. Therefore, the D-vine copula represents the dependence structure of the measurements of solar power sup-

ply and meteorological data sufficiently well with respect to Kendall's rank correlation coefficient.

4.2 Conditional means of solar power supply

In this section we further analyze the dependence structure of solar power supply in the current time period and various other input variables. For this purpose, 250000 realizations were drawn from the D-vine copula shown in Figure 7, using the algorithm explained in [13,1] and its implementation in R, see the VineCopula package in [21]. The conditional empirical means of solar power supply are computed in dependence of GHI, for given lower, middle and upper values of further input variables, see Figure 8. To plot the graphs shown in Figure 8, a partition of the unit interval into 40 parts is used.

Figure 8 visualizes how strongly the conditional means of solar power supply depend on further input variables for given values of GHI in the exemplary time period between 14 and 15 CET. The larger the discrepancy is between the red and blue lines the stronger is the conditional correlation, see [22]. If the red line is above the corresponding blue line in Figure 8 the conditional correlation is positive and otherwise negative. Figure 8 indicates that the conditional means of solar power input are independent of temperature, whereas solar power supply in the previous hourly period and in the hourly period before the previous hourly period are positively correlated with the conditional means of solar power supply. On the other hand, precipitation and humidity are negatively correlated to the conditional means of solar power supply for given values of GHI. Thus, our analysis shows that temperature does not provide additional information to GHI for the probabilistic prediction of solar power supply in the considered time period.

4.3 Validation scores for conditional level-crossing probabilities

To quantify the impact of the considered explanatory variables on the probabilistic prediction of solar power supply conditional level-crossing probabilities are computed by means of the fitted D-vine copula model and compared to corresponding quantities obtained from the measurements of solar power supply using various validation scores. For example, the bias, Brier score (BS), Brier skill score (BSS), empirical correlation coefficient (Corr), reliability (Rel), resolution (Res) and uncertainty (Unc) are considered. Clearly, the bias should be near zero, BS and Rel as low as possible, and BSS and Res as high as possible. Note that BS shows the accuracy of the computed level-crossing probabilities and its information can be decomposed into Rel, Res and Unc. Furthermore, Unc does not depend on the selected model but shows the variability of the measurements. In addition, the considered BSS is used to compare the BS of the D-vine copula model with the BS of the climatological mean. For definitions and further details we refer to [17,27].

4.3.1 Analysis of the goodness-of-fit for different time frames

In this section the goodness of model fit is investigated for different time frames, which are described in Table 4. By comparing different hourly time periods the impact of the solar elevation angle on solar power supply can be determined. In addition, the comparison of the results obtained for different monthly time periods shows the impact of seasonal changes.

Time frame	Description
1	Each hour of day and month separately
2	Each hour of day separately, whereas combining all months
3	Each month separately, whereas combining all hours of day
4	Two consecutive hourly time periods and months
5	Three consecutive hourly time periods and months

Table 4 Time frames for model fitting.

The D-vine copula model fitted for time frame No. 2 in Table 4 leads to the best average scores for bias, BS, BSS, Corr and Rel, whereas time frame No. 5 leads to a slightly better Res than time frame No. 2. Based on the results shown in Table 5, it becomes apparent that fitting the model for each hour of day separately improves the conditional level-crossing probabilities of solar power supply. On the other hand, no further improvement is obtained if each month is considered separately. Thus, the information regarding the solar elevation angle seems to be more important than the information regarding seasonal changes.

Time frame	BS	BSS	Corr	Bias	Rel	Res	Unc
1	0.120	0.495	0.712	0.027	0.004	0.122	0.238
2	0.101	0.575	0.760	0.021	0.003	0.140	0.238
3	0.121	0.493	0.719	0.075	0.011	0.128	0.238
4	0.106	0.553	0.746	0.028	0.003	0.135	0.238
5	0.104	0.564	0.759	0.053	0.007	0.142	0.238

Table 5 Validation scores for conditional level-crossing probabilities of the fitted multivariate D-vine model for different time frames.

4.3.2 Analysis of the goodness-of-fit for different copula models

In this section, the validation scores bias, BS, Brier BSS, Corr, Rel, Res and Unc of the conditional level-crossing probabilities obtained from the multivariate D-vine copula model fitted in Section 4.1 are compared with those obtained from a bivariate Frank copula, which merely models the dependence

between solar power supply and GHI. Note that a Frank copula was proposed as prediction model for solar power supply in [17]. By comparing the scores computed based on both models the impact of further input variables on the probabilistic prediction of solar power supply can be quantified.

Period of time	BS	BSS	Corr	Bias	Rel	Res	Unc
09-10	0.083	-0.146	0.353	0.072	0.031	0.021	0.072
10-11	0.140	0.436	0.673	0.029	0.038	0.148	0.249
11-12	0.111	0.554	0.748	-0.022	0.012	0.151	0.248
12-13	0.104	0.580	0.762	-0.001	0.015	0.158	0.248
13-14	0.109	0.562	0.755	0.028	0.030	0.171	0.249
14-15	0.135	0.452	0.682	0.057	0.017	0.127	0.247
15-16	0.113	-0.012	0.310	0.036	0.028	0.025	0.112
09-16	0.114	0.523	0.726	0.028	0.006	0.131	0.238

Table 6 Validation scores for conditional level-crossing probabilities of the fitted bivariate Frank copula model.

In Tables 6 and 7 the validation scores for conditional level-crossing probabilities are computed for each hourly time period separately and, on the other hand, for the entire time period from 9-16 CET. Note that the scores for the entire time period are usually not the averages over all scores obtained for the hourly time periods. The validation scores show that the multivariate D-vine copula model leads to a considerable improvement compared to the bivariate Frank copula model. Thus, the impact of the additionally considered input variables on the probabilistic prediction of solar power supply is important.

Period of time	BS	BSS	Corr	Bias	Rel	Res	Unc
09-10	0.075	-0.047	0.401	0.056	0.021	0.019	0.072
10-11	0.108	0.566	0.757	0.041	0.010	0.151	0.249
11-12	0.114	0.540	0.735	0.002	0.009	0.145	0.248
12-13	0.101	0.593	0.773	0.007	0.013	0.160	0.248
13-14	0.103	0.585	0.772	0.016	0.017	0.162	0.249
14-15	0.114	0.539	0.742	0.022	0.013	0.146	0.247
15-16	0.093	0.172	0.421	0.006	0.010	0.027	0.112
09-16	0.101	0.575	0.760	0.021	0.003	0.140	0.238

Table 7 Validation scores for conditional level-crossing probabilities of the fitted multivariate D-vine model.

5 Conclusion

Using the D-vine copula model considered in this paper, we determined suitable explanatory variables for the design of probabilistic prediction models for solar power supply at single feed-in points. This knowledge is very important because most probabilistic prediction models considered in the literature have limitations regarding the selection of input variables which can be modelled in a reasonable way. For that purpose, the dependence structure of meteorological input variables, solar power supply measured in previous hourly periods and solar power supply measured for the exemplary time period between 14 and 15 CET was analyzed in detail.

Based on our analysis it turned out that temperature is indeed correlated with solar power supply, but has no additional impact on the probabilistic prediction of solar power supply if considered in combination with GHI. Humidity and precipitation were determined as conditionally negatively correlated to solar power supply in the exemplary time period between 14 and 15 CET for given values of GHI, whereas solar power supply in previous hourly periods seemed to be conditionally positively correlated for given values of GHI. Thus, humidity, precipitation, and solar power supply in the previous hourly period indicate a clear impact on the probabilistic prediction of solar power supply in the current time period.

Moreover, the goodness of model fit was investigated for different time frames and for different copula models. For that purpose, conditional level-crossing probabilities were computed for the exemplary high threshold of 0.7 where various scores were used to validate them. Our analysis showed that the goodness of model fit improves if hours of day are considered separately, whereas considering months separately gives no further improvement. Furthermore, it turned out that considering other input variables, besides GHI, clearly has a positive effect on the accuracy of the computed conditional level-crossing probabilities.

References

1. Aas, K., Czado, C., Frigessi, A., Bakken, H.: Pair-copula constructions of multiple dependence. Insurance: Mathematics and Economics **44**(2), 182–198 (2009)
2. Almeida, M.P., Perpinan, O., Narvarte, L.: Pv power forecast using a nonparametric pv model. Solar Energy **115**, 354–368 (2015)
3. Andrade, J.R., Bessa, R.J.: Improving renewable energy forecasting with a grid of numerical weather predictions. IEEE Transactions on Sustainable Energy **8**(4), 1571–1580 (2017)
4. Balasubramanian, T.N., Appadurai, A.N.: Climate policy. In: V. Venkatramanan, S. Shah, R. Prasad (eds.) Global Climate Change and Environmental Policy, pp. 37–54. Springer (2020)
5. Bayindir, R., Colak, I., Fulli, G., Demirtas, K.: Smart grid technologies and applications. Renewable and Sustainable Energy Reviews **66**, 499–516 (2016)
6. Bessa, R.J.: On the quality of the Gaussian copula for multi-temporal decision-making problems. In: 2016 Power Systems Computation Conference (PSCC), pp. 1–7 (2016)
7. Copernicus Atmosphere Monitoring Service: Open scource global horizontal irradiation data. URL http://www.soda-pro.com/web-services/radiation/cams-radiation-service

8. Golestaneh, F., Gooi, H.B.: Multivariate prediction intervals for photovoltaic power generation. In: 2017 IEEE Innovative Smart Grid Technologies-Asia (ISGT-Asia), pp. 1–5. IEEE (2017)
9. Golestaneh, F., Gooi, H.B., Pinson, P.: Generation and evaluation of space–time trajectories of photovoltaic power. Applied Energy **176**, 80–91 (2016)
10. Haghi, H.V., Lotfifard, S.: Spatiotemporal modeling of wind generation for optimal energy storage sizing. IEEE Transactions on Sustainable Energy **6**(1), 113–121 (2014)
11. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer (2009)
12. Huang, J., Perry, M.: A semi-empirical approach using gradient boosting and k-nearest neighbors regression for GEFCom2014 probabilistic solar power forecasting. International Journal of Forecasting **32**(3), 1081–1086 (2016)
13. Joe, H.: Dependence Modeling with Copulas. Chapman and Hall/CRC (2014)
14. Karimi, M., Mokhlis, H., Naidu, K., Uddin, S., Bakar, A.: Photovoltaic penetration issues and impacts in distribution network - A review. Renewable and Sustainable Energy Reviews **53**, 594–605 (2016)
15. Konishi, S., Kitagawa, G.: Information Criteria and Statistical Modeling. Springer (2008)
16. Leisch, F.: A general framework for finite mixture models and latent glass regression in R. Journal of Statistical Software **11**(8), 1–18 (2004)
17. von Loeper, F., Schaumann, P., de Langlard, M., Hess, R., Bäsmann, R., Schmidt, V.: Probabilistic prediction of solar power supply to distribution networks, using forecasts of global horizontal irradiation. Solar Energy (submitted)
18. Lu, Q., Hu, W., Min, Y., Yuan, F., Gao, Z.: Wind power uncertainty modeling considering spatial dependence based on pair-copula theory. In: PES General Meeting— Conference & Exposition, pp. 1–5. IEEE (2014)
19. Nelsen, R.B.: An Introduction to Copulas. Springer (2006)
20. Papaefthymiou, G., Kurowicka, D.: Using copulas for modeling stochastic dependence in power system uncertainty analysis. IEEE Transactions on Power Systems **24**, 40 – 49 (2009)
21. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2014). URL http://www.R-project.org/
22. Rässler, S.: Statistical Matching: A frequentist Theory, Practical Applications, and alternative Bayesian Approaches. Springer (2012)
23. analysis for Research, M.E.R., 2, A.V.: Open source meteorlogical data. URL https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/
24. SolarPower Europe: Global market outlook 2018-2022 (2017). URL http://www.solarpowereurope.org/wp-content/uploads/2018/09/Global-Market-Outlook-2018-2022.pdf
25. Stadtwerke Ulm/Neu-Ulm Netze GmbH: test areas smart grids. URL https://www.ulm-netze.de/unternehmen/projekt-smart-grids
26. Wan, C., Zhao, J., Song, Y., Xu, Z., Lin, J., Hu, Z.: Photovoltaic and solar power forecasting for smart grid energy management. CSEE Journal of Power and Energy Systems **1**(4), 38–46 (2015)
27. Wilks, D.S.: Statistical Methods in the Atmospheric Sciences. Academic Press (2011)
28. Zhang, B., Dehghanian, P., Kezunovic, M.: Spatial-temporal solar power forecast through use of gaussian conditional random fields. In: IEEE Power and Energy Society General Meeting (PESGM), vol. IEEE, pp. 1–5 (2016)

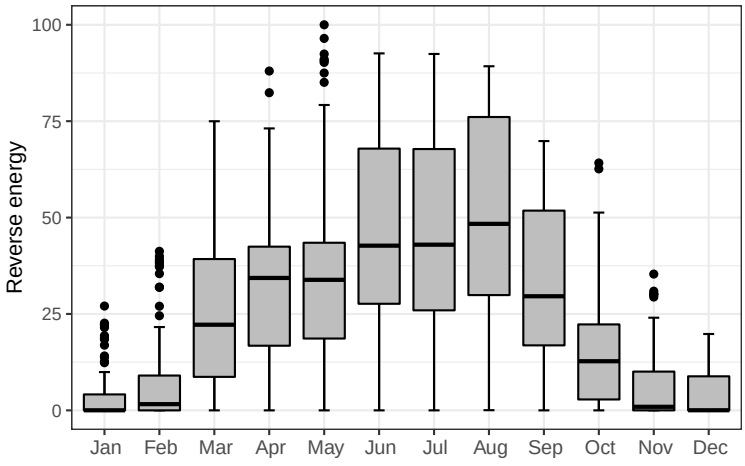

Fig. 1 Daily accumulated reverse energy for each month.

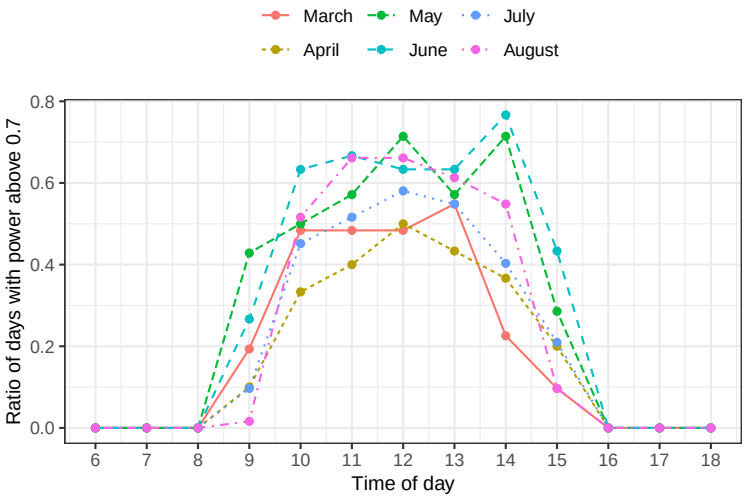

Fig. 2 Ratio of considered days with solar power supply above the threshold of 0.7 for each time of day.

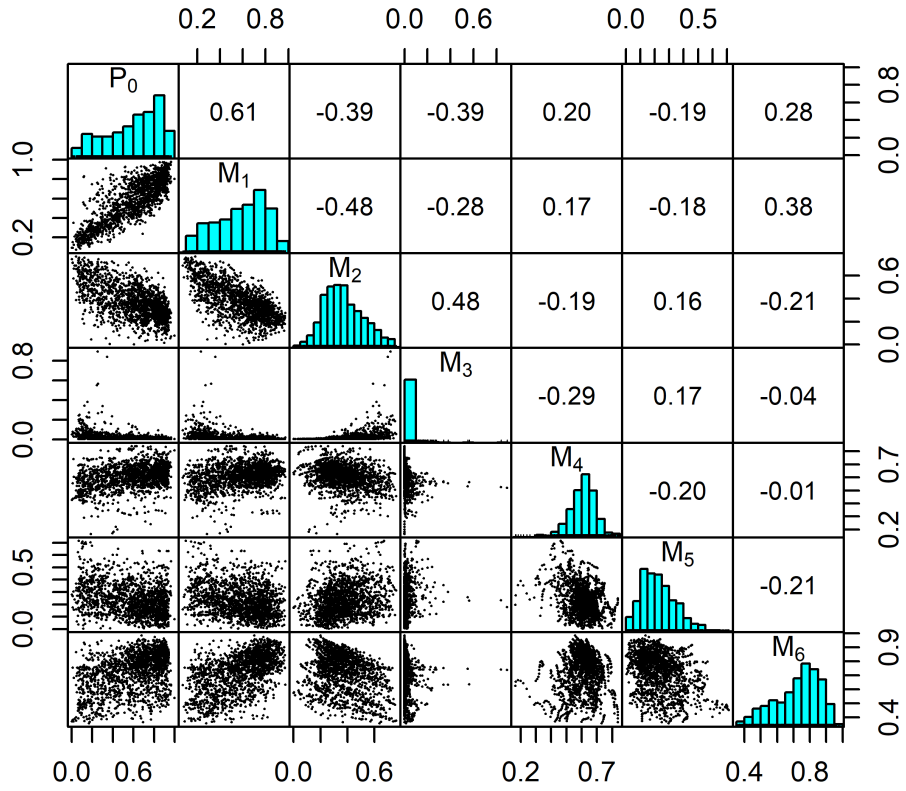

Fig. 3 Pairwise scatterplots, histograms and Kendall's rank correlation coefficients for solar power supply and meteorological variables in the current time period.

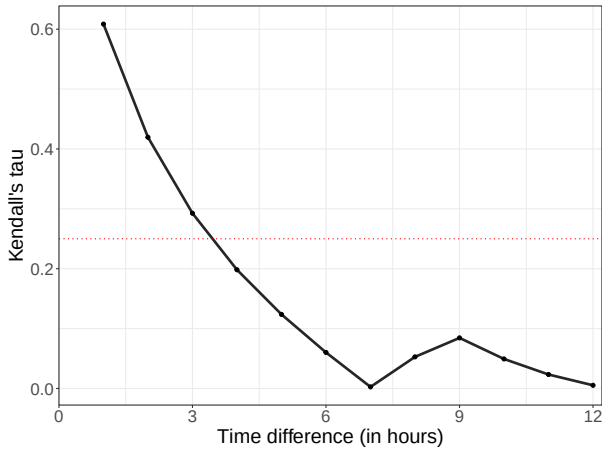

Fig. 4 Kendall's rank correlation coefficients for solar power supply in the current time period of one hour length and solar power supply in previous hourly periods.

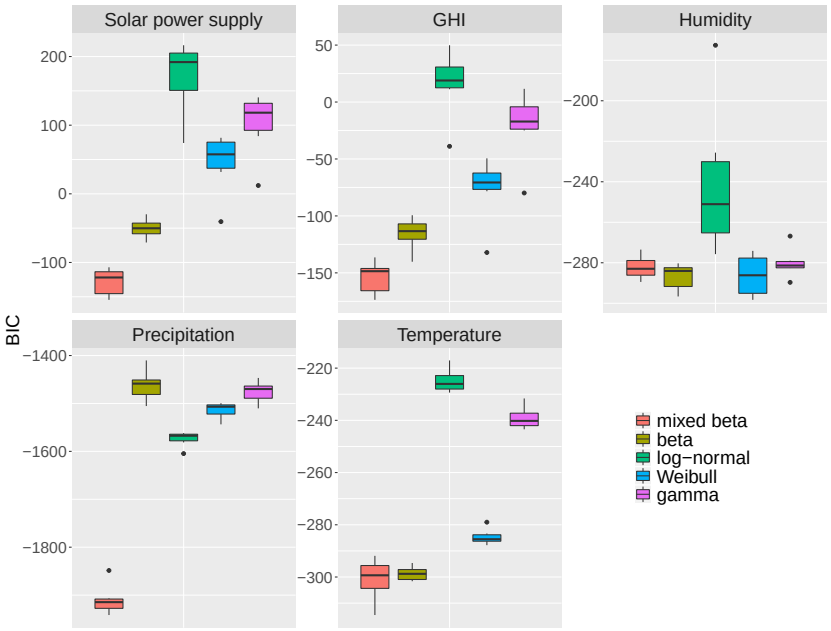

Fig. 5 BIC scores for beta, mixed beta, log-normal, Weibull and gamma distributions.

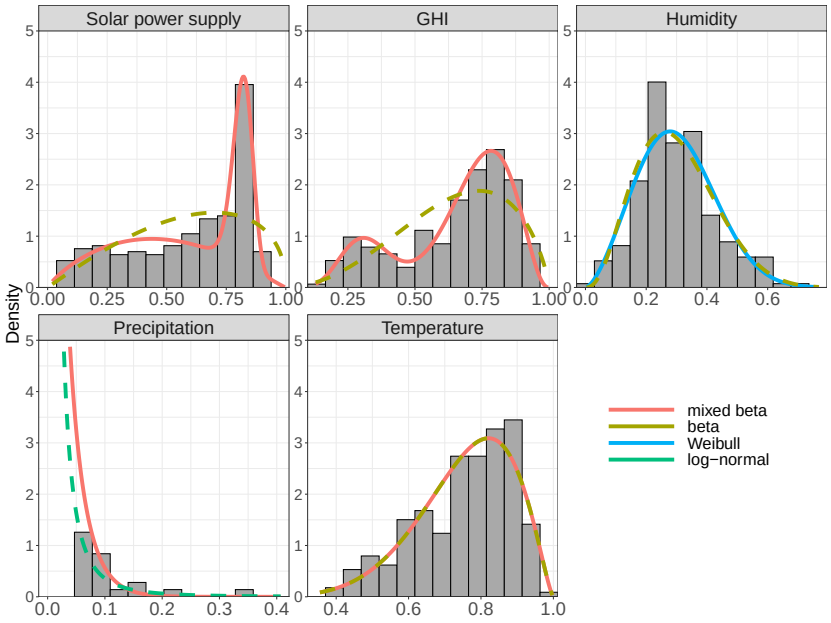

Fig. 6 Fitted marginal densities at the exemplary hour of day 14-15 CET. Solid lines correspond to the best average BIC, whereas dashed lines correspond to the second best average BIC.

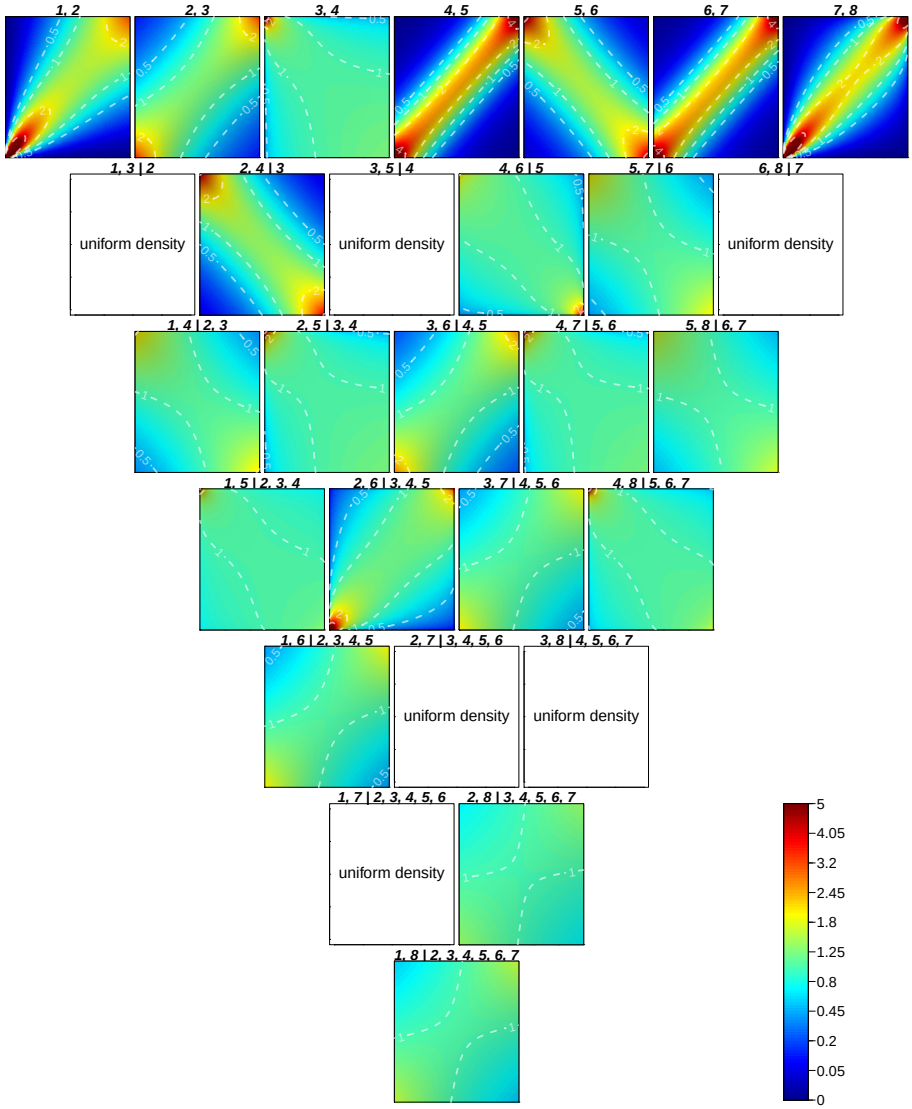

Fig. 7 The D-vine structure fitted to the variables specified in Table 2 for the exemplary hour of day 14-15 CET.

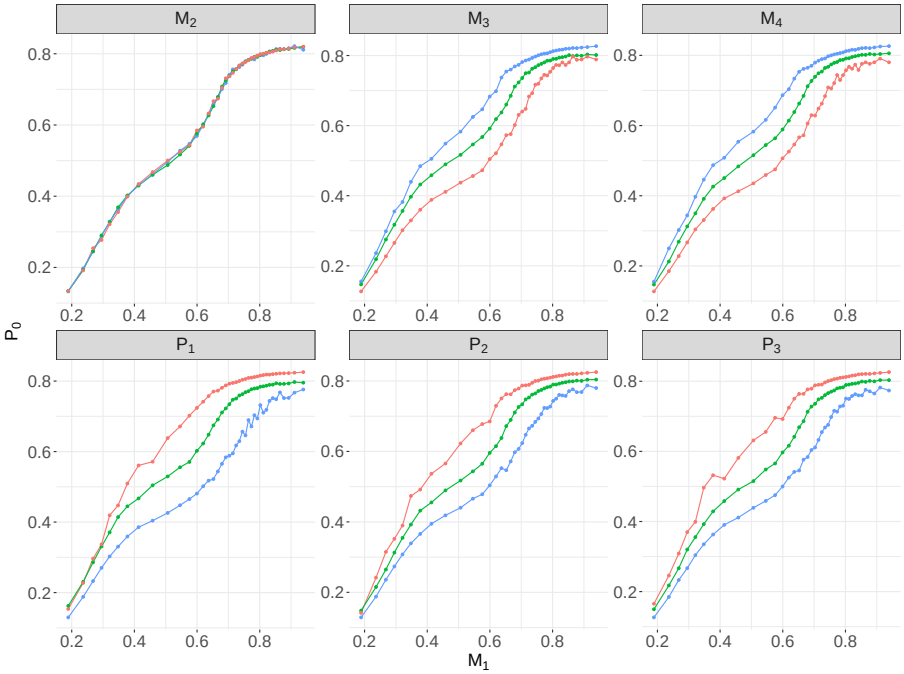

Fig. 8 Conditional empirical means of solar power supply in the current period of one hour length depending on global horizontal irradiation given that the value of a further input variable (specified in the title of the panel) is in the highest 25 percent (red), the middle 50 percent (green), the lowest 25 percent (blue).