Energy 263 (2023) 125610

Contents lists available at ScienceDirect

Energy

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

# A statistical approach to modeling the variability between years in renewable infeed on energy system level

Christopher Jahns *, Paul Osinski, Christoph Weber

*University Duisburg–Essen, Germany*

ARTICLE INFO

*Keywords:*
Renewable energy
Energy system model
Time series
Scenarios

ABSTRACT

Energy system models often rely on assumptions about the infeed of renewable energies. Despite their significance, the renewable time series are often based on single weather years, selected without applying clear criteria. For planning purposes of photovoltaic plants or heating and cooling systems, it is common practice to artificially create weather years composed of months from different years. However, there are only few models for the composition of artificial weather years that represent a well-defined high- or low-infeed-scenario. A new method is proposed to artificially construct infeed time series on system level. Under the assumption of a normal distribution, we compose an infeed time series which aims at meeting a certain quantile of annual infeed. Thus, it is possible to construct different infeed scenarios, to model the inter-year variability of the renewable infeed. The method at hand can be useful for everyone who uses exogenous infeed time series in energy modeling.

## 1. Introduction

Medium and long-term models of the electricity system are based on assumptions about the infeed of renewable energies. As the share of renewable energies increases, these assumptions become more and more important for the results of the models. The first best solution to model a large range of climatic scenarios is to execute full model runs for a set of weather years (normally 20 to 30 [e.g. 1]). Since the execution of system models for 30 weather years implies a high computational effort, it is a common approach in both industry and academia to generate hourly time series for renewable infeed based on a single representative weather year [2–5]. The chosen weather data is then typically used to model the infeed of an expected renewable fleet or the historical infeed time series of the weather year is just scaled with the expected installed capacity of the future year under investigation. Often, the weather year is selected based on expert knowledge. If the inter-year variability is in focus, extreme scenarios are covered by taking a "high" or a "low" weather year into account without further quantification of their probability of occurrence [e.g. 6]. In this paper, a new method is proposed that aims at the construction of an artificial weather year with a given probability, that the renewable infeed exceeds the infeed in the created scenarios. This enables the definition of the infeed scenarios for electricity system models based on objective criteria.

There is scarce literature on artificial weather generation year that addresses interannual variability in the context of large-scale electricity system models. Still, typical hourly weather time series over one year are widely used and discussed for technical applications on a smaller scale, especially for planning purposes of photovoltaic plants or heating and cooling systems. Several methods exist that generate a one-year time series of meteorological parameters like temperature, wind speed or global irradiance, by selecting a typical month from a set of historical years for each calendar month. The selected calendar months are then concatenated into a one year data set, called e.g. typical meteorological year (TMY), test reference year (TRY) or design reference year (DRY). Methods to compute these scenarios with a given probability are not as numerous as for the median scenario.

The new method proposed in this paper aims at the construction of an artificial weather year with a given probability that is composed of twelve real historical weather months. Put differently, for each calendar month, a historical weather month is selected so that the renewable infeed of the composed year corresponds to a prespecified quantile of the distribution of cumulative annual infeed. The basic idea is as follows: Using the assumption of a multivariate normal distribution of the monthly capacity factors of renewable time series, a relation between monthly and yearly probability levels is derived. Based on both, the monthly and yearly probability distributions, an algorithm is designed that minimizes the difference of the yearly capacity factor to the capacity that is related to the target probability level of the individual months.

* Corresponding author.
*E-mail address:* christopher.jahns@uni-due.de (C. Jahns).

https://doi.org/10.1016/j.energy.2022.125610
Received 5 July 2021; Received in revised form 23 September 2022; Accepted 27 September 2022
Available online 1 October 2022
0360-5442/

**List of symbols**

**Symbols & Meaning**

| | |
|---|---|
| $CF^Y$ | Capacity factor for one year (random variable) |
| $CF_m$ | Capacity factor for month m (random variable) |
| $M$ | Set of months Jan, …, Dec |
| $Y$ | A set of years with observed renewable infeed time series |
| $Tech$ | A set of technologies with intermittent infeed e.g. PV |
| $\alpha^Y$ | Target cumulative probability value of $CF^Y$ |
| $\beta_m$ | Target cumulative probability value of $CF_m$ |
| $q_m$ | Target quantile of $CF_m$ |
| $q^Y$ | Target quantile of $CF^Y$ |
| $\Phi$ | Cumulative standard normal distribution function |
| $\sigma_m$ | Standard deviation of residual in month m |
| $\mu_m$ | Estimated mean value of the capacity factor of $CF_m$ |
| $cf_{m,y}$ | Observed capacity factor for month m and year y |
| $cf_y$ | Observed capacity factor for one year y |
| $\beta_{m,y}$ | Estimated cumulative probability value of $cf_{m,y}$ |
| $\alpha_y$ | Estimated cumulative probability value of $cf_y$ |
| $\mathcal{Y}$ | Artificial year $\mathcal{Y}$ resulting from the selection of pairs $(m, y)\forall m \in M$, where $y \in Y$ |
| $cf_{\mathcal{Y}}$ | Capacity factor for $\mathcal{Y}$ |
| $\alpha_{\mathcal{Y}}$ | Estimated cumulative probability value of $cf_{\mathcal{Y}}$ |

Compared to the selection of representative infeed years, the method has two key advantages: The desired probability level is achieved with a high level of precision, even with a small number of observed years and it is simple enough to find common usage in both academic and industrial studies. The usage of observed time series for full months implicitly guarantees that important real-world features of renewable time series (variance, hourly extremes) are preserved and do not have to be statistically simulated. One example where the methodology can be applied (and in an earlier version has already been so) is the annual mid-term scenario forecast used to set up the levy that German electricity consumers used to pay to finance the infeed tariffs of renewable plants in Germany [7].

The remainder of the paper is structured as follows: Firstly, we provide a literature review and elaborate the scientific novelty of the publication. The methodology is explained in detail in Section 3 and is applied to aggregated German photovoltaic and onshore wind time series in Section 4. In Section 5, several checks are performed to test the robustness of our approach under changes to data and the optimization model. A discussion of the results is presented in Section 6 followed by a conclusion in Section 7.

## 2. Literature review and scientific novelty

The following section provides a review of literature associated to the artificial construction of renewable infeed time series that can be used to analyze the inter-annual variability in the context of large-scale electricity market modeling. This section is structured as follows. First, selected publications on the inter-annual variability of renewable energy sources are summarized. Key aspects of the closely related research field of the generation of typical meteorological year (TMY) are then discussed. As this is an important point of the scientific novelty of this paper, we distinguish between methods that can only be used to construct mean weather years and methods that can be used to construct weather years with a predefined probability of exceedance. Based on this, the scientific novelty is outlined.

The inter-annual variability of renewable energy is of great importance for the analysis of financial risk and the security of supply. In the literature, the inter-annual variability is mostly analyzed using long periods of 30 years or more (e.g. [8]). Time series for such a long period are only available for meteorological data, and the renewable energy infeed has to be simulated. This incorporates a possible source of errors. Using real observed infeed time series is preferable, but usually only small periods are available. Limiting factors are the general availability of data and structural changes over time. With a small sample, it is difficult to estimate the probability of exceedance, especially for years with extreme weather conditions. Therefore, a parametric distribution should be considered for a low variance inference. For example, Goren and Taylan [9] assessed the usage of a normal distribution for the annual global horizontal irradiance (GHI) and concluded that a normal distribution should be preferred over an empirical distribution if the considered time series is shorter than ten years. Jung and Schindler [10] estimated and evaluated multiple parametric distributions for the annual wind energy output in Germany to quantify the inter-annual variability. To our knowledge, there is no description in the present literature on how to create renewable infeed scenarios for the analysis of the effect of the inter-annual variability with large-scale electricity market models without relying on time series of 30 years or more.[1] Nevertheless, the generation of weather files, e.g. the TMY, is a closely related research field and the methods and insights may be useful for large-scale electricity market models as well.

Weather data files, such as the TMY (typical meteorological year), are used for the planning of photovoltaic systems, or heating and cooling systems. A TMY is an artificial year with hourly time series of meteorological variables and consists of twelve observed typical meteorological months (TMM) [12]. The aim of TMY methods is to produce a dataset that represents long-term mean weather conditions while retaining statistical properties. There is a wide variety of such methods [12]. Numerous meteorological variables can be considered for a TMY. The methods differ among other things in the weighting and selection of the available variables, which has only a small relevance for the generation of renewable infeed scenarios. Therefore, in this summary we focus on the general method of TMM selection and neglect the different possibilities of weighting schemes and the selection of meteorological variables. TMY methods are, with few exceptions, only suitable to generate a mean meteorological year. Since the scientific novelty of the approach in this paper lies in the generation of weather scenarios with a predefined exceedance probability, we describe only the two most common methods and then go into more detail on methods for the generation of TMYs for extreme weather years afterwards. The Sandia method presented by Hall et al. [13] consists of two stages [14]. First a set of five candidate months for each TMM is chosen using a weighted sum of the distance of the meteorological variables to their long-term cumulative distribution. From these candidate months one month is selected by a measure that among others includes the distance of the minimum, maximum and mean values for temperature and wind speed to their long-term mean. The Danish method is another common approach [15,16]. In a first step, the months are evaluated according to a distance measure, which is based on the distance of the intra-month standard deviation and mean values to their long-term means respectively. Then the maximum of this measure for all meteorological variables is calculated and the three months with smallest maximum are selected as candidates. From these candidates, one month is then selected using a score. Within this score, it is evaluated whether the mean for each meteorological variables is within the one-sigma range around the long-term mean.

[1] A widely discussed topic are methods for the selection of smaller periods (e.g. weeks or days) to reduce the computational burden for market or grid models [e.g. 11]. These approaches tend to lose information compared to a full time series and are not suited to analyze the inter-year variability.

The following is a summary of recent publications on TMY generation improvements. Sun et al. [12] have developed a TMY method specifically for the energy simulation of buildings with daylight use. This method is based on the Sandia method. In addition to changing the focus on daylight harvesting, the method has been adapted by using the Kolmogorov–Smirnov distance and using a genetic algorithm to optimize the weighting scheme. Polo et al. [17] suggested to alter the Sandia method by using the integral of the spectral irradiance and the ratio of the total irradiance of the spectrum to the photon flux density as additional variables. Li et al. [18] and Huang [19] have evaluated different TMY methods with a simulation of the energy consumption of a building. Hosseini et al. [20] proposed to use a machine learning approach for the weight selection to improve the TMM selection. Ernst and Gooday [21] suggested using a TMY with higher resolution, as the hourly resolution can lead to biases when modeling photovoltaic energy yields.

In the following, we provide further insight into methods for creating meteorological or renewable infeed time series with a predefined probability of exceedance. In the solar engineering community, a TMY with a 90 percent exceedance probability is called TMY P90. Mainly commercial tools offer TMY P90 time series, but either select a complete consecutive year closest to a predefined target for the annual value [22] or do not publish the full methodology used to create them [23]. Wey et al. [24] attempt to create the TMY P90 dataset by selecting TMMs that by themselves have an exceedance probability close to 90 percent. They neglected the fact that the probability value of the individual months is not equal to the probability of the composed annual value (as shown in Section 3). Cebecauer and Suri [25] present a method to create a TMY P90 based on the SolarGIS approach. The first step is to calculate a target annual value of the TMY using the assumption of a normal distribution. In a second step, they sort the individual months. They begin with the month with the lowest value as first candidates and iteratively approach the annual target by swapping candidate months with their neighbor in the ranked order until the annual target is reached. They use the annual target as the only measure for the TMM selection. Therefore, the probabilities of exceedance for the individual months are unevenly distributed. The authors argue that a real weather year is composed of individual months with a different probability of occurrence. Still, this might be an disadvantaged. We want to select months close to their (conditional) expected value in order to minimize our error. Therefore, we argue that similar probability values for the different TMMs improve the method.

Fernández Peruchena et al. [26] explain a method to generate a solar irradiation TMY with a probability of exceedance with the purpose of risk assessment of solar power projects. They analyze the annual distribution of GHI and DNI values and concluded that a normal distribution or a Weibull distribution provide a satisfactory fit using a statistical test. Fanego et al. [27] propose a method to construct a TMY with a predefined annual probability of exceedance. In a first step they derive an annual target percentile value using an estimated Weibull distribution. Then they calculate the accommodating target values for the individual months (monthly expected value) with a minimization problem with the constraints that the cumulative monthly targets are equal to the annual target percentile. The objective is a weighted sum of quadratic deviations from the monthly median value multiplied with a weighting factor. The weighting consists of the product of a factor that resembles the months' variability and one for the absolute deviation from their long-term mean respectively. Hence, ceteris paribus the target values for months with higher variability or with a higher absolute value move further away from the median value. Still, while it is plausible that the target value of months with a high variance should be further away from long-term median in absolute values, they do not provide a statistical reasoning for their method. They try to construct a scenario with a predefined probability of exceedance ad hoc and did not try to derive the connection between the monthly and annual probability of exceedance analytically. The publication that, to our knowledge, is closest to the proposed method in this paper is a conference distribution by Jones [28]. Jones [28] make the assumption of a Gaussian distribution and try to derive monthly target values analytically for the case of solar irradiance. However, the approach is not fully described and we are not able to understand and retrace how the formulas have been derived.

In this paper, we present a novel, mathematical consistent method for generating artificial renewable infeed time series consisting of a selection of twelve individual months which is also generalized to the joint creation of multiple infeed time series. The months are selected using a target quantile derived analytically from the given probability of exceedance for the cumulative infeed of the entire year using the assumption of a normal distribution. Additionally, we introduce the constraint that the probability of exceedance is the same for all months. The usage of a parametric distribution in combination with the selection of observed months enables an improvement compared to an empirical selection of whole representative renewable infeed years in cases where we do not have decades of observations in order to validly assess the probability of exceedance. The literature on this topic being scarce, we have only found three methods that are comparable to our method. Our contribution is a scientific novelty as Cebecauer and Suri [25] do not provide an analytical approach for the TMM selection. In contrast to Fanego et al. [27] and Jones [28] we provide a statistically sound method. Compared to Cebecauer and Suri [25] we optimize the selection of the TMMs so that the TMMs have a similar probability of exceedance resulting in a smoother and more representative scenario. To push further we include the possibility of autocorrelation and cross-correlation in our analyses which makes it possible to combine multiple renewable time series and still select months that have a similar probability of exceedance.

## 3. Proposed construction of the reference year

In the following, we describe the construction of a weather year based on real historical weather months of multiple years. To compare the generation of periods with varying installed capacity, we use the capacity factor as a measure for the relative generation of a weather month. We define the capacity factor as $CF = \sum_{t=1}^{T}(P_t/C_t)/T$, where $P_t$ and $C_t$ are the generation and installed capacity and $T$ is the total number of time steps in the considered month. We pursue the objective that solar and onshore wind power generation meet a certain annual quantile $q^Y$. This means that the event that the total production is lower than the total production of our scenario has a predefined probability $\alpha^Y$. Under the assumption of a normally distributed capacity factor, the calculation of $\alpha^Y$ is straightforward, even for small observation samples. For larger samples, the quantile may also be determined using other distributional assumptions or based on observations only.

When composing a weather year based on selected weather months, it should be noted that the selection of extreme months will result in an even more extreme yearly infeed. While in most years, single months with an unusually high weather-induced renewable generation occur, it is much less likely that all months of a certain year show a significantly higher generation relative to the respective months of other years. To account for this phenomenon the relation between the target quantile $q^Y$ for the capacity factor of the whole year and the target quantile $q_m$ of the months is derived in Section 3.1.

To extend the amount of available data to calculate $q_m$, we split the observations of the monthly capacity factor into a deterministic and a stochastic part, to estimate the parameters of the normal distribution. Thereby, we eliminate the obvious dependency between $CF$ and the calendar month (as e.g. solar infeed is higher in summer). Doing so, we can determine the target value of the stochastic part of the monthly capacity factor $CF_m$ based on data of all calendar months, instead of calculating target values for each calendar month individually. The used method is presented in Section 3.2.

The target value $q_m$ is used in the selection process of historical weather months for each calendar month, described in Section 3.3. In addition to the yearly quantile, we aim at individually similar probabilities for all selected months, to avoid the selection of some months with unusually high infeed and some months with unusually low infeed to model an average weather year. Therefore, in the selection process we seek to achieve two goals: To minimize the distance to the target yearly capacity factor and to minimize the distance of each month to the target monthly capacity factor. Thus, a linear mixed-integer problem (MIP) is proposed.

### 3.1. Converting quantiles for the monthly and yearly infeed

This section aims to show the conversion of a targeted quantile for a capacity factor for a whole year into target quantiles for the specific months. We start with the assumption that the capacity factors $CF_m$ for the individual months are independently normally distributed:

$$CF_m \sim \mathcal{N}(\mu_m, \sigma_m^2) \quad \forall\ m \in M, \tag{1}$$

where M is a set of calendar months from January to December. The capacity factor of one year $CF_y$ equals the weighted mean of the capacity factors of the individual months with a weighting factor $w_m$ based on the different number of days.

$$CF^Y = \sum_{m \in M} w_m CF_m, \quad where \sum_{m \in M} w_m = 1. \tag{2}$$

Since $CF^Y$ is the weighted mean of an independently normally distributed random variable, it is also normally distributed.

$$CF^Y \sim \mathcal{N}(\sum_{m \in M} w_m \mu_m, \sum_{m \in M} w_m^2 \sigma_m^2). \tag{3}$$

We start the construction with a target probability level $\alpha^Y$ and quantile $q^Y$, so that $Pr[CF_y < q_y] = \alpha^Y$. Hence, we are in control of the probability of the constructed weather year. To select individual months, we need the corresponding probabilities $\beta_m\ \forall\ m \in M$ to construct the quantiles $q_m$. Under the assumption of an independent normal distribution, the quantile $q^Y$ can be rewritten as follows.

$$q^Y = \sum_{m \in M} w_m \mu_m + \Phi^{-1}(\alpha^Y) \sqrt{\sum_{m \in M} w_m^2 \sigma_m^2}. \tag{4}$$

$$q_m = \mu_m + \Phi^{-1}(\beta_m)\sigma_m \quad \forall \quad m \in M. \tag{5}$$

Inserting $q^Y$ and $q_m$ for the corresponding $CF$, Eq. (2) leads to

$$q^Y = \sum_{m \in M} w_m q_m. \tag{6}$$

We get

$$\Phi^{-1}(\alpha^Y) \sqrt{\sum_{m \in M} w_m^2 \sigma_m^2} + \sum_{m \in M} w_m \mu_m = \sum_{m \in M} w_m \Phi^{-1}(\beta_m)\sigma_m + \sum_{m \in M} w_m \mu_m. \tag{7}$$

If we use the additional constraint that $\beta_m$ is the same for every $m \in M$ to ensure that the infeed of the months are equally likely, we can derive the relation between the target value of $\alpha^Y$ and the target value for the individual months $\beta_m$.

$$\beta_m = \Phi\left(\Phi^{-1}(\alpha^Y) \frac{\sqrt{\sum_{m \in M} w_m^2 \sigma_m^2}}{\sum_{m \in M} w_m \sigma_m}\right). \tag{8}$$

$\beta_m$ can be converted to a target value $q_m$, which is later used to select a weather month per calendar month.

When taking correlation between the different months into consideration, we have to define the preceding in a more general way using vector notation and a multivariate normal distribution:

$$CF^Y = \mathbf{w}^T \mathbf{CF}_m \sim N\left(\mathbf{w}^T \boldsymbol{\mu}, \mathbf{w}^T \Sigma \mathbf{w}\right), \tag{9}$$

where $\Sigma$ is the variance–covariance matrix of the vector $\mathbf{CF}_m$.

Analogous to Eq. (7), we get

$$\Phi^{-1}(\alpha^Y)\sqrt{\mathbf{w}^T \Sigma \mathbf{w}} + \mathbf{w}^T \boldsymbol{\mu} = \sum_{m \in M} w_m \Phi^{-1}(\beta_m)\sigma_m + \mathbf{w}^T \boldsymbol{\mu}, \tag{10}$$

and

$$\beta_m = \Phi\left(\Phi^{-1}(\alpha^Y) \frac{\sqrt{\mathbf{w}^T \Sigma \mathbf{w}}}{\sum_{m \in M} w_m \sigma_m}\right). \tag{11}$$

For now, we have only considered a single time series. However, multiple time series, e.g. onshore wind and photovoltaic (PV) infeed, can be considered with a simple adaption of the preceding: The index $m$ in Eq. (11) is replaced by the indices $m$ and $\tau$, where $\tau$ is an element of a set of the technologies.

Similar to Eq. (2), the total capacity factor for all technologies for year $y$ is then defined as

$$CF^Y = \sum_{m \in M, \tau \in Tech} CF_{m,\tau} w_{m,\tau}, \tag{12}$$

where $CF_{m,\tau}$ is the capacity factor for month $m$ and technology $\tau$ and $w_{m,\tau}$ is a weighting factor that accounts for the different number of days in different months and the different capacities of the technologies. $\mathbf{CF}_{m,\tau}$ is defined as the stacked vector of all technologies' monthly capacity factors.

$$\mathbf{CF}_{m,\tau} = \begin{bmatrix} CF_{Jan,wind} \\ \vdots \\ CF_{Dec,wind} \\ CF_{Jan,PV} \\ \vdots \\ CF_{Dec,PV} \end{bmatrix}. \tag{13}$$

Based on this notation, the method for a single time series may be extended in a straightforward way. This leads to

$$\beta_{m,\tau} = \Phi\left(\Phi^{-1}(\alpha^Y) \frac{\sqrt{\mathbf{w}^T \Sigma \mathbf{w}}}{\sum_{m,\tau} w_{m,\tau}\sigma_{m,\tau}}\right), \tag{14}$$

where $\Sigma$ is the variance–covariance matrix of the vector $\mathbf{CF}_{m,\tau}$, $\sigma_{m,\tau}$ is the standard deviation of the residuals of month m and technology $\tau$.

### 3.2. Estimating the conditional mean and the variance–covariance matrix

To calculate $q_m$ and select the months of the weather years, we have to estimate the deterministic part $\mu$ and the stochastic part $\Sigma$ of the capacity factors. Generally, many different methods are suited to estimate $\mu$, including e.g. taking the conditional mean. We suggest the use of a regression model with a Fourier series of second order. This model has the advantage that it includes the a priori knowledge of a sinusoidal seasonal pattern and therefore has less parameters to be estimated. This will lead to more robust regression results if only limited data are available. The regression model is shown in Eq. (15).

$$\begin{aligned} \hat{\mu}_{m,t} = b_0 + b_1 sin(\frac{2}{12}\pi m_t) + \\ b_2 cos(\frac{2}{12}\pi x) + b_3 sin(\frac{4}{12}\pi m_t) + b_4 sin(\frac{4}{12}\pi m_t), \end{aligned} \tag{15}$$

where $m_t$ is a number from one to twelve indicating the calendar month and $\hat{\mu}_t$ is the predicted value of the expected monthly capacity factor. For the estimation of the variance–covariance matrix, the sample variance–covariance matrix can be used as an estimator. Still, some covariances should be zero and we use this information to minimize the number of parameters to be estimated. For example, the autocorrelation is only reasonable up to a certain lag and it might be more efficient to only allow a covariance between two infeed time series in the same month.

### 3.3. Selecting the weather months using an mixed-integer optimization model

With Eq. (11) or (14) we can calculate the target quantiles $q_m$ for the capacity factors of the months. We can create an artificially composed infeed time series by chosing combinations $(m, y, \tau)\ \forall\ m \in M$ and $\tau \in Tech,$ where $y \in Y$ while trying to match our predefined quantiles. For example, we may choose the infeed time series of January 2001, February 2010, etc. In general, there will not be any observed months with an exact match to the target quantile. As explained above, we aim at the simultaneous minimization of (a) the distance between the capacity factor of the selected calendar months $cf_m$ and the target quantile $q_m$ and (b) the distance between the resulting yearly capacity factor and the target quantile $q_y$. As the variance differs greatly between different months and technologies, it is reasonable to look at a distance that is independent of the variance. We suggest using the corresponding probability values for the observations, which can be calculated as follows.

$$\beta_{m,y,\tau} = \Phi(\frac{cf_{m,y,\tau} - \mu_{m,y,\tau}}{\sigma_{m,\tau}})\ . \tag{16}$$

Hence, (a) can be written as the minimization problem,

$$\min \sum_{(m,y,\tau)\in\mathcal{Y}} |\beta_{m,y,\tau} - \beta_m|, \tag{17}$$

where $\mathcal{Y}$ is a set of the selected combinations $(m, y, \tau)\ \forall\ m \in M$ and $\tau \in Tech,$ where $y \in Y$, for the artificial constructed infeed year. With $\alpha_{\mathcal{Y}}$ being calculated as

$$\alpha_{\mathcal{Y}} = \Phi\left(\frac{\sum_{(m,y,\tau)\in\mathcal{Y}} w_m \Phi^{-1}(\beta_{m,y,\tau})\sigma_{m,\tau}}{\sqrt{\mathbf{w}^T \Sigma \mathbf{w}}}\right), \tag{18}$$

objective (b) can be formulated as:

$$\min \quad |\alpha^{\mathcal{Y}} - \alpha^Y|. \tag{19}$$

For the selection, a multi-criteria optimization problem can be used. It is furthermore useful to preserve the relationship between the infeed of different technologies. Therefore, we suggest applying the additional constraint that the same year and month is used for every technology. To make the choice of a set of appropriate weightings less difficult, we allow a specific range for objective (b), while minimizing objective (a). The problem can be defined as follows:

$$\begin{aligned}
\underset{\lambda_{m,y}}{\text{minimize}} \quad & \sum_{m\in M,\tau\in Tech,y\in Years} (\lambda_{m,y}|\beta_{m,y,\tau} - \beta_m|) \\
\text{subject to} \quad & \sum_{m\in M,\tau\in Tech} \lambda_{m,\tau}\Phi^{-1}(p^Y_{m,\tau})\sigma_{m,\tau} < \sqrt{w^T \Sigma w}(\alpha^Y + \epsilon) \\
& \sum_{m\in M,\tau\in Tech} \lambda_{m,\tau}\Phi^{-1}(p^Y_{m,\tau})\sigma_{m,\tau} > \sqrt{w^T \Sigma w}(\alpha^Y - \epsilon) \\
& \sum_{y\in Years} \lambda_{m,y} = 1 \quad \forall\ m \in M,
\end{aligned}$$

where $\lambda_{m,Y}$ are binary variables. The first and second constraints ensure that $\alpha^Y$ is within the predefined tolerance. The third constraint states that only one year can be selected for every month.

## 4. Application to simulated German onshore wind and PV time series

As a proof of concept, the presented method is applied to German onshore wind and PV infeed time series, whereby the weather years are constructed for the target values $\alpha^Y$ of 0.1 and 0.9. We use hourly time series of capacity factors in Germany from renewables.ninja (RN) [29, 30] as our main dataset. This dataset is available for weather years between 1985 to 2018 and the capacity factors are simulated based on the power plant fleet of 2016.

We use the regression model of Section 3.2 to estimate $\mu$. In Table 1 the regression results are given. The models include four Fourier terms derived from the calendar month wind and a constant. The standard errors of the coefficients are given in parentheses. The coefficient of determination $R^2$ indicates that 96 percent of the variance of the PV capacity factor and 47 percent of the onshore capacity factor can be explained with this model. Additionally, the Durbin–Watson statistic, the Breusch–Pagan statistics and, as the assumption of a normal distribution is used, the results of the Shapiro–Wilk test are given for each model. The values of 1.875 for PV and 1.895 for onshore wind in terms of the Durbin–Watson statistic indicate a small positive autocorrelation. The hypothesis of an autocorrelation of zero can be rejected for PV at a confidence level of 0.1. The Breusch–Pagan statistics show that a significant heteroscedasticity exists in the residuals. The statistics in Table 1 show that the null hypothesis of the Shapiro–Wilk test, that the residuals are normally distributed, can only be rejected for the onshore wind time series. The distribution is still close to a normal distribution and an approximation is reasonable. Fig. 1 shows the fit of the sinusoidal regression fit. The heteroscedasticity conditional on the calendar month can be observed visually, especially for the onshore wind time series.

While estimating the variance–covariance matrix, it can be advantageous to define additional constraints. These constraints should be as restrictive as necessary to reduce the variance of the estimation, while an overly restrictive constraint could eliminate important features. Possible assumptions are homoscedasticity ($\sigma_m = const.\quad \forall m \in M$) or an autocorrelation of zero. We do not use the assumption of homoscedasticity, because of the visual difference in variance (Fig. 1) and the results of the Breusch–Pagan test (Table 1). Regarding autocorrelation, the Durbin–Watson statistics show that the hypothesis of an autocorrelation of zero at the lag of one cannot be rejected for onshore wind whereas it is expected to be small for PV. Hence, the assumption of an autocorrelation of zero is applied subsequently. For the relation between the infeed of the technologies, we allow the cross-correlation in the same calendar month. Hence, a correlation between the January PV and January onshore wind is considered, while a correlation between the January PV and February onshore wind is disregarded. Using Eq. (11), we can compute $\beta_m$. For a scenario with $\alpha^Y = 0.1$ we obtain a $\beta_m$ of 0.38, for a scenario with $\alpha^Y = 0.9$ the $\beta_m$ is 0.62.

After the estimation of $\mu$ and $\Sigma$, we use the optimization model described in Section 3.3. We allow for a deviation $\epsilon$ from $\alpha^Y$ of 0.005. Fig. 2 shows the capacity factors of all historical weather months taken into account for each calendar month. The months selected by the optimization algorithm are marked in the respective color. The corresponding years of the selected months are indicated in Table A.1 in the Appendix. The first thing to be noticed is that, even though two extreme scenarios regarding the yearly infeed have been constructed, the selected months are very close to each other. This illustrates the fallacy when extreme months are selected to construct a weather year with a high total infeed. The selected months for onshore wind exemplify how the observed heteroscedasticity is handled in the model. During the winter months, when the variance of the capacity factor is high, the selected months for the 10 percent and 90 percent scenario are further apart.

## 5. Quality and robustness checks

Following the illustrated example of the application of the proposed method, we perform four quality and robustness tests. While we aim at taking reasonable decisions along the process, we evaluate how different options can influence the results. The first robustness check is the use of different constraints on the estimation of $\Sigma$. In a second test, we try to use heuristics to select the years, since using this approach reduces the implementation effort. After that, we evaluate if the use of historically observed time series might lead to a different result, compared to simulated time series. In a final robustness check, we alter the period for the estimation, to get a sense of how many

**Table 1**
Regression results for the monthly capacity factors.

| | *Dependent variable:* | |
|---|---|---|
| | PV | Onshore |
| $\sin(\frac{2\pi}{12}$ Month $)$ | $-0.009^{***}$ | $0.029^{***}$ |
| | (0.001) | (0.004) |
| $\cos(\frac{2\pi}{12}$ Month$)$ | $-0.083^{***}$ | $0.065^{***}$ |
| | (0.001) | (0.004) |
| $\sin(\frac{4\pi}{12}$ Month$)$ | $-0.001$ | $0.010^{***}$ |
| | (0.001) | (0.004) |
| $\cos(\frac{4\pi}{12}$ Month$)$ | $-0.008^{***}$ | 0.0004 |
| | (0.001) | (0.004) |
| Constant | $0.125^{***}$ | $0.193^{***}$ |
| | (0.001) | (0.003) |
| DW test P Value | $1.875^{*}$ | 1.895 |
| BP Test | $35.389^{***}$ | $64.643^{***}$ |
| SW Test | 0.996 | $0.964^{***}$ |
| Observations | 408 | 408 |
| $R^2$ | 0.966 | 0.473 |
| F Statistic (df = 4; 403) | $2{,}829.352^{***}$ | $90.341^{***}$ |

*Note:* $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

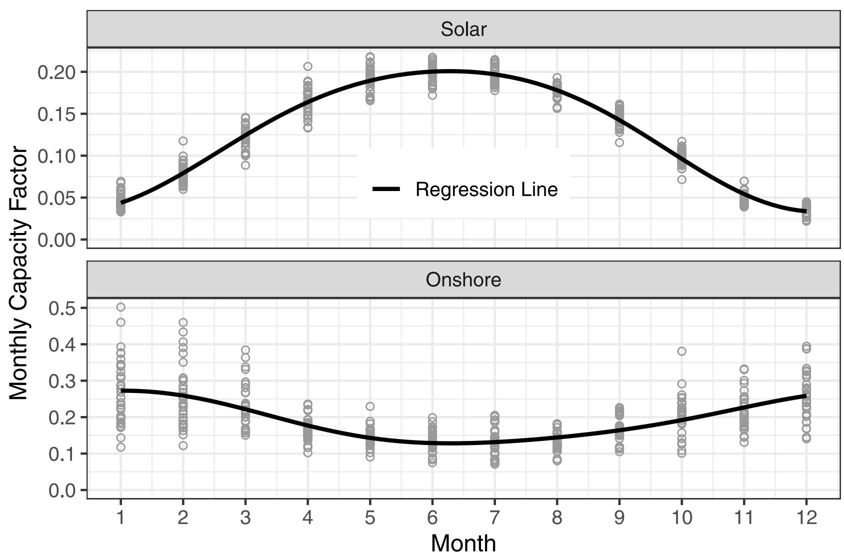


**Fig. 1.** Capacity factors of onshore wind and PV with corresponding regression line.

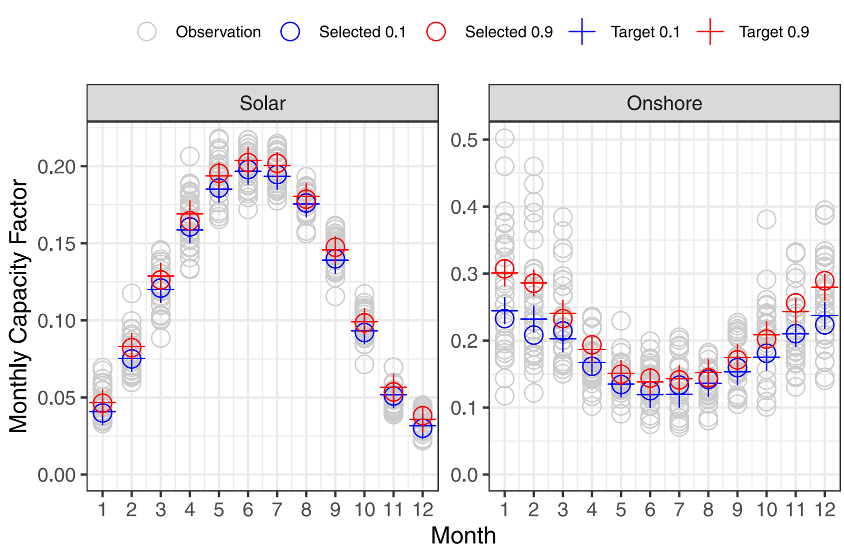


**Fig. 2.** Composed year and month selection.

**Table 2**
Robustness checks concerning the estimation of the variance–covariance matrix.

| Case | Target $\alpha^{Y}$ | Recalculated $\alpha^{Y(R)}$ |
|---|---|---|
| Full | 0.1 | 0.083 |
| Full | 0.9 | 0.917 |
| Diagonal | 0.1 | 0.085 |
| Diagonal | 0.9 | 0.915 |

years one might need for a reliable estimation. The results are shown in Table 3. The first column $\alpha^{Y}$ represents the initial target for the probability value. In column $\alpha_{\mathcal{Y}}$, the observed probability from the artificially created year is shown. The base case is taken from Section 4. We can observe that the realized probability value deviates from our target within the given tolerance. As an alternative measure, the target capacity factor $CF^{Y}$ and the realized capacity factor $CF_{\mathcal{Y}}$ are given. As additional information, the average absolute deviation from our monthly target $\beta_m$ is given.

### 5.1. Altering the constraints while estimating the variance–covariance matrix

In Section 4 we describe how the use of additional constraints for the estimation of the variance–covariance matrix can lead to a more efficient estimation. In the following, we evaluate the robustness to different constraints in estimating the variance–covariance matrix. For this purpose, we use the relation between the probability value $\beta_m$ and $\alpha^{Y}$, given in Eq. (14). In a first step, we estimate the variance–covariance matrix with alternative constraints and calculate a corresponding $\beta_m^{A}$ for a target probability value $\alpha^{Y}$. In the second step, we use the estimated variance–covariance matrix $\Sigma$ from Section 4, $\beta_m^{A}$ and recalculate the probability value for $CF^{Y}$ ($\alpha^{Y(R)}$) as a measure of how robust our choices are against misspecification of constraints on the variance–covariance matrix. Table 2 shows the results of this robustness check. Based on the fully estimated variance–covariance matrix, without constraints, we would miss our initial target by roughly two percentage points compared to Section 4. If we use a diagonal matrix, neither allowing auto- nor cross-correlation, it affects our initial target similarly. The method is thus relatively robust to other options in the estimation, but we still suggest using reasonable constraints.

### 5.2. Quality loss through multidimensional approximation

Using the proposed method with multiple infeed time series separately is an alternative approach envisaged. Yet it implies inconsistencies measured by the total infeed. With this robustness test, we assess the error we make if we choose individual time series for the 90 and 10 percent scenario. Hence, we choose months creating a 90 percent case, once using only onshore wind and once using only PV infeed. Using the selected months, and the estimated $\Sigma_{m,t}$ from Section 4 with the multivariate approach, we can calculate how big the error would be if we had chosen to look at only the single time series separately. While we miss the target $\alpha^{Y}$ by roughly five percentage points (Table 3, Case "Multi Time series") we miss the target capacity factor only by a small amount. By looking at the two time series separately, we lose precision in terms of the probability of occurrence, but we only make a small error in terms of absolute values. Measured by the MAE $\beta_m$ we even get better performance. This is due to the additional degree of freedom compared to the base scenario, where the restriction that the same month should be used for PV and onshore wind was necessary to keep the cross-correlation. However, depending on the exact object of research, there are good reasons to make sure the same months are selected for both time series.

### 5.3. Using heuristics of months with capacity factors lying next to the target

Using the MIP solver for the weather month selection, which is described in Section 3.3, is the part that requires the most effort in implementing the proposed method. Therefore the following heuristic is proposed. We reduce the set of possible months from which we can select to the months whose capacity factors in terms of Euclidean distance are adjacent to the target $CF_m$. Thus we have $2^{12} = 4096$ different combinations to choose from. While we ensure that the deviation from $\beta_m$ is small, the problem becomes simple enough to iterate through all possible solutions. We then can use the solution with the smallest deviation to our target $\alpha^{Y}$. This heuristic is evaluated in Table 3 (case "Heuristic"). As we do not include a pre-specified tolerance level, this heuristic is coming closer to the target $\alpha^{Y}$, but the performance is slightly worse, measured by the mean absolute deviation of $\beta_m$. In summary, this heuristic is a valid alternative to solving an MIP.

### 5.4. Historical or simulated data

In Section 4 and in the prior subsections of Section 5 we used simulated time series from RN. We choose this option because using observed infeed time series for the composition of statistically representative time series based on real weather months has one major issue. The capacity factors from two different years are expected to differ due to technological progress,[2] especially regarding onshore wind power, even if the weather was the same in both years. However, the selection of statistically representative months should be based solely on weather conditions.

For the modeling of energy systems, some may still prefer real observed infeed time series to avoid errors arising from the simulation process of renewable infeed. Therefore, we conduct a robustness check using a dataset of hourly capacity factors from open-power-system-data.org (OSD), which is based on the publications by German TSOs and grid regulators.

The OSD time series are obtained from 2012 to 2018. In a first step, we apply our original method to a subset of the RN data from the same years to receive a selection of months $\mathcal{Y}^{\mathcal{RN}}$ for both the 0.1 and 0.9 case. Then, we estimate $\Sigma_{OSD}$ and $\mu_{OSD}$ using the OSD dataset, but when calculating $\alpha^{Y}$ we use the months selected based on $\mathcal{Y}^{\mathcal{RN}}$. The results are shown in 5.5. We end up at a quantile of 0.058 for the 0.1 scenario and of 0.794 for the 0.9 scenario. So the target is missed by a greater margin, but still a high and a low infeed scenario are reproduced correctly. However, in terms of the monthly prediction, the method lacks robustness as the value of 0.19 for the MAE of $\beta_m$ is almost three times higher than in the base scenario. Because there is only a limited systematic bias these errors cancel each other out, resulting in an only moderate deviation regarding $\alpha^{Y}$.

### 5.5. Altering the length of the time series

The simulated time series, used in Section 4, covers a period of more than 30 years. This should be sufficient for a reliable estimation of the means and of the variance–covariance matrix. However, a sufficient period might not always be available. Therefore, we evaluate the robustness to changes in the length of the time series. Using the RN dataset from Section 4 but only a period from 2012 to 2018, we select historical weather months for the 10 and 90 percent cases. Based on these months, we evaluate the error relative to the results of the full period. Table 3 shows that again the deviation in terms of $\alpha^{Y}$ is large, but small in terms of the absolute capacity factor. Hence, similar to the check of a different dataset (Section 5.4), the method can be used based on shorter periods to produce a high and low infeed scenario with relatively large deviations regarding monthly quantiles.

[2] And less importantly also due to differences in the regional distribution of plants.

**Table 3**
Results quality and robustness checks.

| | Cum. Probability | | | Capacity factor | |
|---|---|---|---|---|---|
| | Year Target $\alpha^Y$ | Year Observed $\alpha_y$ | Month Deviation MAE $\beta_m$ | Year Target $CF^Y$ | Year Observed $CF_y$ |
| Base (Section 4) | 0.1 | 0.096 | 0.052 | 0.150 | 0.150 |
| Base (Section 4) | 0.9 | 0.904 | 0.060 | 0.169 | 0.169 |
| Multi time series (Section 5.2) | 0.1 | 0.056 | 0.032 | 0.150 | 0.147 |
| Multi time series (Section 5.2) | 0.9 | 0.950 | 0.034 | 0.169 | 0.171 |
| Heuristic (Section 5.3) | 0.1 | 0.100 | 0.063 | 0.150 | 0.150 |
| Heuristic (Section 5.3) | 0.9 | 0.900 | 0.077 | 0.169 | 0.169 |
| Different data (Section 5.4) | 0.1 | 0.058 | 0.138 | 0.138 | 0.136 |
| Different data (Section 5.4) | 0.9 | 0.794 | 0.194 | 0.154 | 0.151 |
| Different years (Section 5.5) | 0.1 | 0.062 | 0.124 | 0.149 | 0.147 |
| Different years (Section 5.5) | 0.9 | 0.775 | 0.175 | 0.169 | 0.163 |

## 6. Discussion

Enabling the generation of a weather year with a quantifiable probability of occurrence, our method is clearly advantageous compared to the usage of single weather years for extreme scenarios of energy system simulation. Due to its comparatively low computational effort, it offers an alternative to full energy system simulations with multiple weather years. Adapting methods that are used to generate typical meteorological years (TMY) in PV planning applications, our procedure is a statistically profound way to generate TMY P90 and P10 cases for energy system simulations.

Compared to Cebecauer and Suri [25], the presented procedure allows dealing simultaneously with infeed time series of various technologies. Furthermore, it differs in one key aspect, that should be beneficial in most applications. Cebecauer and Suri [25] propose one method for constructing a median case (P50 scenario) and another for lower probabilities. Both rely on the assumption of a normal distribution of the yearly values. In contrast to Cebecauer and Suri [25], we explicitly require at an evenly distributed probability of occurrences throughout the months. Cebecauer and Suri [25] construct a TMY P50 scenario by minimizing for each month the difference to a median hourly duration curve and the corresponding average value, to ensure a yearly distribution matching the multiannual distribution. As no similar property may be exploited for other cases (e.g. the so-called TMY P90), Cebecauer and Suri [25] construct a year with high variability. They argue that a real year with low infeed experiences a high variability of the residuals. Even though this is true, it is unlikely that the extreme months from the artificial year created by the method of Cebecauer and Suri [25] are concurrent to an arbitrary observed year. It is much more likely that they will produce larger errors, whereas we try to minimize the monthly deviations by using moderate probability values. Additionally, with the proposed method, it is also possible to minimize the distance to a corresponding duration curve, as we can use shorter periods than a month for the weather year construction. The disadvantage is, that we would have shorter periods of historical observations and thereby more possible structural breaks in the time series.

While the method seems to be quite robust, assessed against different estimations of $\Sigma$ or assessed against a heuristic to select the weather months, the outcome can still change significantly, if a measured instead of a simulated time series of capacity factors or only a subset of years is used for the estimation. Inherently, this is due to the property, that small changes in $\beta_m$ can result in large changes of $\alpha^Y$. This results in small differences in terms of absolute values, but larger deviations in terms of the probabilities of occurrence. However, these differences are somewhat expected. The capacity factors of measured time series do not give an unbiased measure of the weather month, as they are distorted by changes of the power plant fleet. Also a period of seven years is not considered a sufficient statistical basis to model the climatic conditions. Therefore, the observed deviation in terms of monthly and yearly quantiles, while still correctly representing a high and low scenario is not a counter-argument against the validity of the method.

## 7. Conclusion

The selection of scenarios for the infeed of renewables is often approached qualitatively by a rather arbitrary choice of weather years. We present a new way of constructing a renewable infeed scenario for one year designed to meet a targeted quantile for the total infeed. Composing a weather year based on weather months from different years, our approach is similar to the Sandia method [13] and its successors, which are commonly used for planning purposes of single PV projects. This implies the preservation of real-world features of renewable time series like variance or hourly extremes. Compared to the few of those methods that create weather years for target quantiles different from 50 percent (e.g. the TMY P10/P90 in Cebecauer and Suri [25]), our procedure is more direct. Based on the assumption that the infeed of one calendar month in different years can be described by a normal distribution, we derive a relation between yearly and monthly target quantiles and then select historical months, solving a combined minimization problem of the difference to the yearly and monthly targets. Thus, we receive a robust scenario with similar probabilities for the infeed for the different calendar months, making sure all months are as close as possible to the derived monthly target quantile. The method is further adapted to be suitable for multiple time series, e.g. wind and PV.

In a case study, we apply our method to simulated hourly time series of the aggregated infeed of PV and onshore wind in Germany over approx. 30 years. The historical months, that were thus selected (cf. Appendix, Table A.1) can be used by further researchers that want to create P10 and P90 cases for Germany, especially, but not exclusively, when using the same data set [29,30]. Robustness checks showed, that if measured historical instead of simulated infeed data or a shorter period are used, the procedure is still able to reproduce a high or low scenario, but the monthly and yearly target quantiles are not met exactly. Further sensitivity checks, using a heuristic instead of the optimization to select the best historical months or applying the method to single time series in a framework of multiple renewable technologies, proved the method to be robust towards these changes. Using a heuristic instead of the mixed-integer problem for the selection of the months can be reasonable for easier implementation. The error of using the method applied to multiple time series is small, but still, should be avoided. We advise using this method for scenario building for studies on energy system level, as it is more quantitative and reasonable than the selection of single weather years but still more lightweight than a full energy system simulation of multiple years.

**Table A.1**
Selected weather months.

| $\alpha^Y$ | 0.1 | 0.5 | 0.9 |
|---|---|---|---|
| Jan | 2004 | 2014 | 2000 |
| Feb | 2010 | 2017 | 2014 |
| Mar | 1997 | 2004 | 2004 |
| Apr | 2002 | 1994 | 2017 |
| May | 1991 | 2003 | 1999 |
| Jun | 1986 | 2015 | 1993 |
| Jul | 2004 | 2004 | 1985 |
| Aug | 1985 | 2001 | 2001 |
| Sep | 2013 | 1986 | 2015 |
| Oct | 1993 | 2012 | 2010 |
| Nov | 1988 | 1986 | 1986 |
| Dec | 2005 | 2012 | 2006 |

## CRediT authorship contribution statement

**Christopher Jahns:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Software. **Paul Osinski:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Christoph Weber:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data will be made available on request.

## Appendix

See Table A.1.

## References

[1] ENTSO-E. Mid-term adequacy forecast 2019 appendix 2, methodology. Technical report, 2019, URL https://eepublicdownloads.blob.core.windows.net/public-cdn-container/clean-documents/sdc-documents/MAF/2019/MAF%202019%20Appendix%202%20-%20Methodology.pdf.

[2] Caglayan DG, Heinrichs HU, Linssen J, Robinius M, Stolten D. Impact of different weather years on the design of hydrogen supply pathways for transport needs. Int J Hydrogen Energy 2019;47:25442–56.

[3] Maclaurin G, Cole W, Lopez A, Reimers A, Rosenlieb E, Roberts B. Understanding inter-annual variability of PV energy production in the contiguous United States. In: 2018 IEEE international conference on probabilistic methods applied to power systems. PMAPS, 2018, p. 1–4.

[4] Brouwer A, Van den Broek M, Zappa W, Turkenburg W, Faaij A. Least-cost options for integrating intermittent renewables in low-carbon power systems. Appl Energy 2016;161:48–74.

[5] 50Hertz Transmission GmbH, Amprion GmbH, TenneT TSO GmbH, TransnetBW GmbH. Szenariorahmen für den Netzentwicklungsplan Strom 2030 (version 2019). Technical report, 2018, URL https://www.netzentwicklungsplan.de/sites/default/files/paragraphs-files/%C3%9CNB-Entwurf_Szenariorahmen_2030_V2019_0_0.pdf.

[6] Zappa W, Jungiger M, van den Broek M. Is a 100% renewable European power system feasible by 2050? Appl Energy 2019;233–234:1027–50.

[7] Baginski P, Bellenbaum J, Beran P, Broll R, Felling T, Jahns C, et al. Mittelfristprognose zur deutschlandweiten stromerzeugung aus EEG geförderten kraftwerken für die kalenderjahre 2019 bis 2023. Technical report, Lehrstuhl für Energiewirtschaft, University Duisburg-Essen; 2018.

[8] Collins S, Deane P, Ó Gallachóir B, Pfenninger S, Staffell I. Impacts of inter-annual wind and solar variations on the European power system. Joule 2018;2(10):2076–90. http://dx.doi.org/10.1016/j.joule.2018.06.020.

[9] Goren D, Taylan O. P50/P90 analysis of a solar photovoltaic plant in METU NCC using the empirical method. In: 2020 2nd international conference on photovoltaic science and technologies. IEEE; 2020, p. 1–8. http://dx.doi.org/10.1109/PVCon51547.2020.9757770.

[10] Jung C, Schindler D. On the inter-annual variability of wind energy generation – a case study from Germany. Appl Energy 2018;230:845–54. http://dx.doi.org/10.1016/j.apenergy.2018.09.019.

[11] Nahmmacher P, Schmid E, Hirth L, Knopf B. Carpe diem: A novel approach to select representative days for long-term power system modeling. Energy 2016;112:430–42.

[12] Sun J, Li Z, Xiao F, Xiao J. Generation of typical meteorological year for integrated climate based daylight modeling and building energy simulation. Renew Energy 2020;160:721–9. http://dx.doi.org/10.1016/j.renene.2020.07.024.

[13] Hall IJ, Prairie RR, Anderson HE, Boes EC. Generation of a typical meteorological year. Technical report Report SAND-78-1096C, CONF-780639-1, Sandia Laboratories; 1978.

[14] Gazela M, Mathioulakis E. A new method for typical weather data selection to evaluate long-term performance of solar energy systems. Sol Energy 2001;70(4):339–48. http://dx.doi.org/10.1016/S0038-092X(00)00151-1.

[15] Lund H. The Design Reference Year. 1995.

[16] Zang H, Xu Q, Wang W, Chen K. A step-by-step application of Danish method for generation of typical meteorological years of China. J Renew Sustain Energy 2013;5(5):053145. http://dx.doi.org/10.1063/1.4826193.

[17] Polo J, Alonso-Abella M, Martan-Chivelet N, Alonso-Montesinos J, Lapez G, Marzo A, et al. Typical meteorological year methodologies applied to solar spectral irradiance for PV applications. Energy 2020;190:116453. http://dx.doi.org/10.1016/j.energy.2019.116453, URL http://www.sciencedirect.com/science/article/pii/S0360544219321486.

[18] Li H, Yang Y, Lv K, Liu J, Yang L. Compare several methods of select typical meteorological year for building energy simulation in China. Energy 2020;209:118465. http://dx.doi.org/10.1016/j.energy.2020.118465.

[19] Huang K-T. Identifying a suitable hourly solar diffuse fraction model to generate the typical meteorological year for building energy simulation application. Renew Energy 2020;157:1102–15. http://dx.doi.org/10.1016/j.renene.2020.05.094.

[20] Hosseini M, Bigtashi A, Lee B. A systematic approach in constructing typical meteorological year weather files using machine learning. Energy Build 2020;226:110375. http://dx.doi.org/10.1016/j.enbuild.2020.110375.

[21] Ernst M, Gooday J. Methodology for generating high time resolution typical meteorological year data for accurate photovoltaic energy yield modelling. Sol Energy 2019;189:299–306. http://dx.doi.org/10.1016/j.solener.2019.07.069.

[22] Thomas C, Wey E, Blanc P. Typical meteorological year generation service methodology report. Technical report, MINES ParisTech / ARMINES / Transvalor; 2017, URL http://www.soda-pro.com/documents/10157/2948104/Transvalor_TMY_generation.pdf/e2393d5f-587a-4a67-b2f8-8083e71c7173.

[23] Vaisala. Vaisala 3TIER services global solar dataset / methodology and validation. Technical report, 2017, Ref. B211641EN-B URL "https://www.vaisala.com/sites/default/files/documents/3TIER%20Solar%20Dataset%20Methodology%20and%20Validation.pdf".

[24] Wey E, Thomas C, Blanc P, Espinar B, Mouadine M, Bouhamidi MH, et al. A fusion method for creating sub-hourly DNI-based TMY from long-term satellite-based and short-termground-based irradiation data. In: SolarPACES 2012. (USBKey Paper 22983). 2012.

[25] Cebecauer T, Suri M. Typical meteorological year data: SolarGIS approach. Energy Procedia 2015;69:1958–69. http://dx.doi.org/10.1016/j.egypro.2015.03.195.

[26] Fernández Peruchena CM, Ramírez L, Silva-Pérez MA, Lara V, Bermejo D, Gastón M, et al. A statistical characterization of the long-term solar resource: Towards risk assessment for solar power projects. Sol Energy 2016;123:29–39. http://dx.doi.org/10.1016/j.solener.2015.10.051.

[27] Fanego VL, Rubio JP, Peruchena CMF, Romeo MG, Tejera SM, Santigosa LR, et al. A novel procedure for generating solar irradiance TSYs. In: AIP conference proceedings, Author(s); 2017, 140015. http://dx.doi.org/10.1063/1.4984523.

[28] Jones RK. Improved methodology for typical meteorological year month selection matching annual irradiance. In: 2021 IEEE 48th photovoltaic specialists conference. IEEE; 2021, p. 0018–22. http://dx.doi.org/10.1109/PVSC43889.2021.9518562.

[29] Staffell I, Pfenninger S. Using bias-corrected reanalysis to simulate current and future wind power output. Energy 2016;114:1224–39.

[30] Pfenninger S, Staffell I. Long-term patterns of European PV output using 30 years of validated hourly reanalysis and satellite data. Energy 2016;114:1251–65.