Chapter 3 Established Mathematical Approaches for Synthetic Solar Irradiance Data Generation

CHAPTER 3 ESTABLISHED MATHEMATICAL APPROACHES FOR SYNTHETIC SOLAR IRRADIANCE DATA GENERATION

Joakim Munkhammar and Joakim Widén

Munkhammar, J. and Widén, J. (2021), “Established mathematical approaches for synthetic solar irradiance data generation,” in Bright, J. M. (ed.), *Synthetic Solar Irradiance: Modeling Solar Data*, Melville, New York: AIP Publishing, pp. 3-1–3-36

3.1 INTRODUCTION

Scarcity of solar irradiance data for different resolutions and for various locations is the essential motivator for the development of synthetic solar irradiance data generators. Such generators are typically based on some more or less elaborate mathematical framework, aiming to represent and reproduce both the systematic and stochastic variations that can be observed in actual irradiance data. Many different mathematical approaches have been developed, evaluated, and applied over the last half century to provide physical and statistical models of solar irradiance, as well as methods to analyze and transform solar irradiance data so that suitable mathematical models can be fitted.

The aim of this chapter is to give the reader a brief introduction to, and overview of, these mathematical approaches, along with examples of their use in the scientific literature on synthetic irradiance modeling. By the end of this chapter, the reader should be able to identify available state-of-the-art synthetic solar irradiance models; classify these into temporal, spatial, and spatiotemporal types; understand the main mathematical concepts behind them; and know how to apply them for synthetic solar irradiance data generation.

https://doi.org/10.1063/9780735421820_003

The chapter is organized as follows. In Sec. 3.2, an overview of state-of-the-art scientific research on methods for synthetic generation of irradiance data is provided, showing the development of the research field from the 1970s until today. This is followed in Sec. 3.3 by a review of different methods to achieve stationarity in solar irradiance time series, which is a necessary step toward being able to model the variability in irradiance as a stochastic process. Section 3.4 introduces probability distribution models for solar irradiance, which form the basis, or at least are a major component, of many synthetic generator models. In the subsequent sections, the major types of established mathematical models for synthetic solar irradiance are discussed. Autoregressive and similar models are treated in Sec. 3.5, Markov chain models in Sec. 3.6, copulas and other multivariate distribution model approaches in Sec. 3.7, and various machine learning and artificial intelligence (AI) methods in Sec. 3.8. A summary is finally included in Sec. 3.9.

3.2 OVERVIEW OF EXISTING APPROACHES

We carried out a literature review to identify the most significant contributions to the field of synthetic solar irradiance modeling. The literature is vast if one considers all papers that are of some relevance for synthetic generation. For example, numerous studies have been made on the characterization of irradiance probability distributions for different time resolutions and sites, as well as analysis of correlation structures in irradiance time series. More recently, the field of solar forecasting has grown substantially, and many of the methods and models in that field could also be relevant in synthetic solar irradiance modeling. However, covering this whole field would not be feasible nor desirable. Rather, we want to isolate the most important examples of models specifically developed for synthetic irradiance generation. Therefore, when choosing the scientific studies to include in this overview, the following limitations were applied:

- Only papers in peer-reviewed scientific journals were considered (that is, not conferences or reports), in order to identify contributions thoroughly put under scrutiny and of high novelty (at their time of publication).
- Papers had to explicitly have the objective of synthetic data generation (not just limited to, e.g., data analysis or model identification).
- Only studies on solar irradiance were included (not, e.g., photovoltaic power generation).
- The papers had to, as judged by us, make significant contributions to the development of the field at the time of publication (merely incremental improvements of existing methods were not included).

Tables 3.1 and 3.2 list the identified studies sorted according to publication date. The tables present the main characteristics of the approaches. Table 3.1 includes studies limited to the temporal dimension, i.e., generation of synthetic time series for individual locations. Table 3.2 includes studies that have approached the more complex problem of generating spatiotemporal data, i.e., spatially

Table 3.1

Methods for synthetic generation of solar irradiance time-series (temporal models only), sorted according to publication date.

Paper	Year	Temporal resolution	Synthetically generated quantity	Generative method
Exell (1976)	1976	Hour	Clearness index	First-order Markov chain model
Brinkworth (1977)	1977	Daily	GHI	First-order AR process
Mustacchi *et al.* (1979)	1979	Hourly	GHI	Autoregressive moving average (ARMA) and Markov chain
Bertoli *et al.* (1981)	1981	Daily	Clearness index	First-order AR process
Exell (1981)	1981	Daily	Clearness index	Second-order random process
Amato *et al.* (1985)	1985	Daily	Clearness index	First-order AR process
Vergara-Dominguez *et al.* (1985)	1985	Daily	Clear-sky index	DARMA process
Balouktsis and Tsalides (1986)	1986	Hourly	Transformed GHI	Fourier transform spectral representation method
Aguiar *et al.* (1988)	1988	Daily	Clearness index	Markov chain
Graham *et al.* (1988)	1988	Daily	Transformed clearness index	First-order AR process
Graham and Hollands (1990)	1990	Hourly	Transformed clearness index	First-order AR process
Aguiar and Collares-Pereira (1992)	1992	Hourly	Normalized clearness index	First-order AR process
Mora-Lopez and Sidrach-de-Cardoba (1998)	1997	Hourly	Normalized clear-sky index	Multiplicative ARMA process
Mohandes *et al.* (1998)	1998	Daily	GHI	ANN
Morf (1998)	1998	Any	Stochastic insolation function (SIM)	Markov chain
Poggi *et al.* (2000)	2000	Hourly	Clearness index	Markov chain
Glasbey (2001)	2001	30 s	Square root of ratio between actual and expected GHI	Nonlinear AR process
Hontoria *et al.* (2001)	2001	Hourly	Clearness index	Neural network
Mora-Lopez and Sidrach-de-Cardoba (2003)	2003	Hourly	Clearness index	Probabilistic finite automata
Mellit *et al.* (2005)	2008	Monthly	Clearness index	Hybrid ANN and Markov chain
Brabec *et al.* (2013)	2013	15 s	SSN and SSSN	Logistic regression
Morf (2013)	2013	Any	SIM and cloud cover	Markov chain

Table 3.1 (*Continued.*)

Paper	Year	Temporal resolution	Synthetically generated quantity	Generative method
Ngoko *et al.* (2014)	2014	1 min	Normalized clearness index	Second-order Markov chain
Bright *et al.* (2015)	2015	1 min	GHI	Markov chain
Munkhammar and Widén (2017a)	2017	1 min	Clear-sky index	Copula
Grantham *et al.* (2017)	2017	5 min	Coupled GHI and DNI clear-sky index	Non-parametric bootstrapping
Grantham *et al.* (2018)	2018	Daily and hourly	GHI	Fourier series and non-parametric bootstrapping
Munkhammar and Widén (2018a; 2018b)	2018	1 min	Clear-sky index	Markov chain mixture
Peruchena *et al.* (2018)	2018	1 min	Coupled GHI and DNI	*K*-medoids clustering and dynamic paths
Zhang *et al.* (2018a; 2018b)	2018	1 min	GHI	Linear interpolation and Cholesky decomposition
Frimane *et al.* (2019)	2019	1 min	GHI	Non-parametric Bayesian clustering and DPGMM
Shepero *et al.* (2019)	2019	1 min	Clear-sky index	Hidden Markov model

correlated time series. This difference between temporal and spatiotemporal modeling is important. That is, temporal modeling implies synthetic irradiance modeling for single locations over time. Spatiotemporal modeling considers the solar irradiance over time at a set of spatially dispersed locations rather than at a single location. Conceptually, were you to use a temporal-only model to produce data at two neighboring sites, there would be no correspondence between the two time series produced (e.g., a cloud passing first over site 1 and then over site 2), except if, in the case of downscaling, the starting data should capture this spatial correspondence already.

Table 3.2

Methods for synthetic generation of spatiotemporal solar irradiance data, sorted according to publication date.

Paper	Year	Temporal resolution	Spatial resolution	Generative method
Cai and Aliprantis (2013)	2013	Flexible	Flexible	Fractal cloud shadow patterns moved over an area
Bright *et al.* (2017)	2017	1 min	Flexible	Circular cloud shapes moved over an area
Jazayeri *et al.* (2017)	2017	Flexible	Flexible	Sequences of real sky images
Munkhammar and Widén (2019)	2019	1 min	Flexible	Markov chain mixture distribution model

For the temporal models in Table 3.1, which make up the bulk of the literature, the temporal resolution that the methods were developed for is specified alongside the synthetically generated quantity (irradiance, clearness index, etc.) and the method used for this synthetic modeling. The quantity listed is the one that is modeled as a (usually) stationary, stochastic process, and the generative method is the stochastic model used for this purpose. The vast majority of the models separate, in one way or another, systematic and periodic variability in the irradiance from stochastic fluctuations. The periodic trends are usually fairly straightforward to model, while the identification and modeling of the stochastic part is more challenging and is usually the main contribution in these studies.

A number of observations can be made from Table 3.1. One thing to note is that the temporal resolution of the synthetic data has become successively higher, starting with daily values in the first papers and developing through to sub-hourly resolution in the most recent studies. This reflects the increasing availability of higher-resolution measurement data at more locations and the improving computing and data storage capabilities, but also the fact that the main intended application in earlier studies was performance evaluation of solar collectors, which does not require very high resolution, whereas the main application recently has been photovoltaic (PV) system simulation, where the response to irradiance variability is instantaneous.

Another thing to note is that the main method used to obtain stationary quantities was the clearness index, which is based on extraterrestrial irradiance, whereas the clear-sky index has been the main choice in more recent studies. Most likely, this also reflects improving computing and data management capabilities, and the successive improvement of clear-sky irradiance models (Sun *et al.*, 2019; Sun *et al.*, 2021). The benefits and drawbacks of using the clearness and clear-sky indices are discussed in Sec. 3.3.1.

Regarding the generative methods for producing synthetic time series, it is also possible to see a significant development, with early studies focusing mainly on autoregressive processes and later studies exploring a wide variety of different model types, tending toward more and more complex approaches.

The very latest development in the field of synthetic irradiance, sparked by the development of grid-integrated, spatially dispersed PV systems, is the venturing into the spatiotemporal domain. For these approaches, the most important of which are listed in Table 3.2, the generative method is typically more complex than the ones in Table 3.1 and not possible to formulate as explicit mathematical formulas.

In the subsequent sections, the methods by which the stationary quantity is obtained are first described, and thereafter the various methods and models by which this quantity is modeled and synthetically generated are treated in more detail.

3.3 METHODS TO ACHIEVE STATIONARITY

A stationary process is a stochastic process whose statistical properties do not change over time. Formally, for a stochastic process $\{X_t\}$ to be strictly stationary, this implies that the probabilistic behavior of a set of values $\{x_1, x_2, \ldots, x_t\}$ is identical to that of a time-shifted set $\{x_{1+\tau}, x_{2+\tau}, \ldots, x_{t+\tau}\}$ (Shumway and Stoffer, 2011, p. 22). That is, if the process is shifted in time, the joint probability distribution of any subset of random values in the process does not change. However, this is an impractical assumption for many purposes. Instead, the notion of a weakly stationary time series can be more useful. Formally, a weakly stationary time series is a finite variance process such that (i) the mean value is constant and does not depend on time t, and (ii) the autocovariance only depends on the separation of points, not time t (Shumway and Stoffer, 2011).

Clearly, solar irradiance is not a stationary process. It exhibits systematic changes in both mean and variance over the course of the day and over the year, due to the sun's apparent motion across the sky. However, many approaches to model stochastic processes and to generate synthetic time series require an assumption on stationarity. This is also the case with most methods for generating synthetic solar irradiance data. Consequently, methods to achieve stationarity by removing systematic trends in solar irradiance data are necessary and, conversely, so are methods for adding these trends to synthetic data generated from stationary processes. Stationarity as a concept will become clearer in Sec. 3.3.1 and with Fig. 3.1.

The main approaches to achieving stationarity in existing synthetic solar irradiance generators are described next. Sometimes several of these methods may be used in combination.

3.3.1 Clearness and clear-sky index

The most commonly used method for removing systematic trends in irradiance data is to normalize it either by the extraterrestrial irradiance or the clear-sky irradiance, obtained from physical models. In the first case, the *clearness index* is calculated, defined as

$$k(t) = \frac{G(t)}{G_0(t)}, \tag{3.1}$$

where G_0 is the extraterrestrial irradiance. The extraterrestrial irradiance is, as the name implies, the global horizontal irradiance (GHI) outside the earth's atmosphere, and is determined directly from the solar constant and the incidence angle of irradiance at time t (Widén and Munkhammar, 2019a).

The clearness index removes the major periodicities in the GHI due to solar angles, but it does not compensate for the attenuation of irradiance that occurs in the atmosphere, which increases with the atmospheric path length and, thus, also with the incidence angle. To additionally remove this systematic trend, the *clear-sky index* can be used instead, defined as

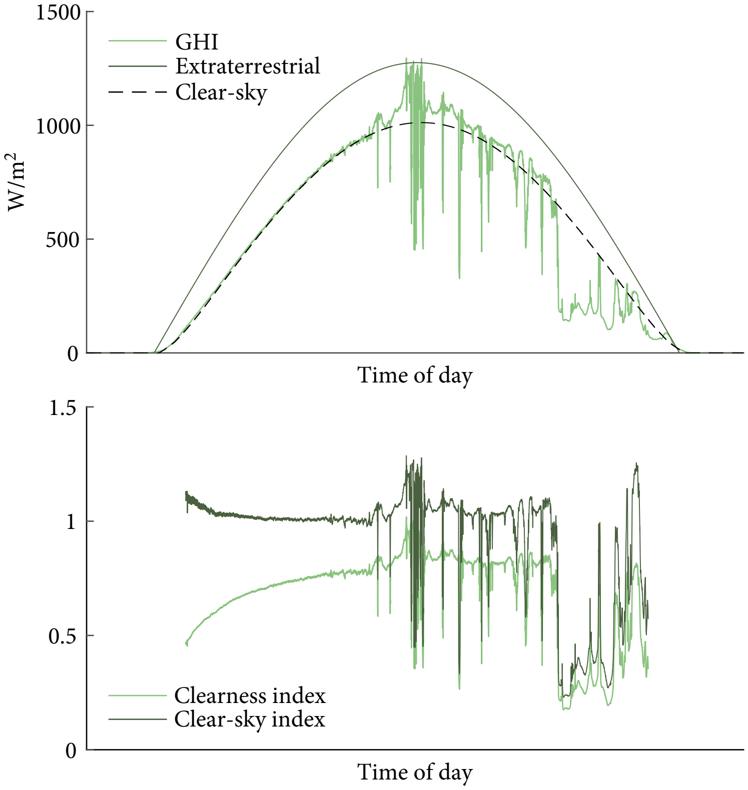

FIG. 3.1

Examples of the clearness and clear-sky index over one day with a 1-second resolution. The upper figure shows GHI data along with modeled extraterrestrial and clear-sky irradiance. The GHI data are from the Oahu Solar Measurement Grid from 25 March 2010 (Sengupta and Andreas, 2020) and the clear-sky irradiance generated with the McClear model (CAMS McClear Service, 2020). The lower figure shows the corresponding indices for solar elevations above 10°.

$$\kappa(t) = \frac{G(t)}{G_c(t)}, \tag{3.2}$$

where G_c is the clear-sky irradiance, i.e., the GHI under clear-sky conditions (no clouds). This removes trends in the GHI caused by both solar angles and atmospheric attenuation, while leaving (theoretically) only stochastic fluctuations due to clouds. Estimating the clear-sky irradiance requires more complex modeling than the extraterrestrial irradiance, but, along with the development of clear-sky models, using the clear-sky index rather than the clearness index has become more popular both in models for synthetic irradiance (cf. Tables 3.1 and 3.2) and in studies of solar irradiance variability. For an overview of different clear-sky models, see Sun *et al.* (2019) and Sun *et al.* (2021).

Figure 3.1 shows examples of the clearness and clear-sky index over the course of one day, along with the radiation components that they are based on: measured GHI, modeled extraterrestrial radiation, and modeled clear-sky irradiance. As can be clearly seen, the extraterrestrial radiation does not remove all of the systematic diurnal variability, resulting in a clearness index that increases during the morning hours. On the other hand, the clear-sky index is very stable during the same time period, due to the modeled clear-sky radiation accurately reproducing the observed radiation.

A benefit of using the clearness index is that it is very simple to obtain as the fraction of GHI and extraterrestrial irradiance, and that the latter to high accuracy is possible to estimate without observational data. A drawback of the clearness index is, as seen in Fig. 3.1, that it does not capture all systematic variations since it does not take into consideration air mass and the variability in atmospheric turbidity. For achieving stationarity, the clear-sky index is therefore advantageous.

3.3.2 Fourier series

Another approach to achieve stationarity is to use Fourier series to identify, model, and compensate for periodicities in irradiance. A Fourier series can be used to approximate periodic functions and is defined, for the function $f(t)$, as (Vretblad, 2003)

$$s(t) = \frac{a_0}{2} + \sum_{n=1}^{N}\left(a_n \cos\left(\frac{2\pi nt}{T}\right) + b_n \sin\left(\frac{2\pi nt}{T}\right)\right), \tag{3.3}$$

where T is the period of the function and the so-called Fourier coefficients are

$$a_n = \frac{2}{T}\int_0^T f(t)\cos\frac{2\pi nt}{T}\,dt,$$
$$b_n = \frac{2}{T}\int_0^T f(t)\sin\frac{2\pi nt}{T}\,dt.$$

Fourier series have been used in several previous synthetic irradiance generators, mainly among the earlier studies. Vergara-Dominguez *et al.* (1985) used an approach similar to the one in Sec. 3.3.1, estimating daily clear-sky irradiance by fitting a Fourier series with t as the day of the year and $T = 365$ to low-pass filtered daily GHI data and then dividing the original time series by this estimated periodic component. The same approach was then used to add the periodic component to synthetic daily data generated with their stationary stochastic process.

Balouktsis and Tsalides (1986) used a similar approach for hourly data, fitting one Fourier series for each hour of the day over the 365 days of the year, but then removing the periodicities by subtracting the periodic components from the original data, rather than dividing by them.

Aguiar *et al.* (1988) chose to use the clearness index for removing seasonal periodicity in daily data, but also compared this approach to Fourier series and a moving average, finding that both the Fourier series approach and the moving average approaches removed the periodic trend to a higher degree than the simpler clearness index approach.

A more recent study that also used the Fourier series approach is Grantham *et al.* (2018), which is well detailed in Chap. 2 (see Eq. 2.10), where separate daily and hourly GHI models were used, each including a periodic seasonal component modeled as a Fourier series with significant frequencies identified using power spectrum analysis.

Figure 3.2 shows daily GHI data over 594 days along with fitted Fourier series, similar to the aforementioned studies. It can be seen that most of the seasonal variation in the irradiance data is captured by the first term of the Fourier series, having an annual period. Adding further terms with shorter periods does not alter the series much. In contrast to the clearness and clear-sky indices, the Fourier series does not have a direct physical interpretation, but is a statistical trend that fits the data in a least-squares sense. However, it can still be as effective in achieving stationarity.

3.3.3 Partitioning into time intervals and bins

If a stochastic process exhibits periodicities or other trends such that the process is approximately stationary or at least is subject to less complex trends on certain time intervals, performing data analysis and modeling on these intervals separately can be an alternative or a complement to the approaches discussed earlier. For example, over the year the solar irradiance has clear seasonal trends, as was seen in Fig. 3.2, but over a month the trend in daily irradiation may not be very pronounced or complex. Consequently, we could study each monthly interval separately to find a simpler trend, such as a linear trend instead of the periodic Fourier series, or even assume that the data are stationary within each monthly interval to avoid trend removal completely.

Using this approach, the very early study by Brinkworth (1977) divided the year into six separate time intervals on which the seasonal periodicity in daily GHI could be simplified into linear trends. Bertoli *et al.* (1981) and Amato *et al.* (1985) calculated the clearness index to first remove the major seasonal periodicities in daily data, then analyzed daily irradiance time series separately for each month of the year to account for differences between monthly clearness index distributions. Poggi *et al.* (2000) also found sufficient stationarity of the clearness index within each month.

Similarly, Balouktsis and Tsalides (1986), after subtracting periodicities modeled with Fourier series, as mentioned earlier, noted that their hourly time series were not yet stationary due to different frequency distributions between different seasons. However, instead of completely dividing the analysis

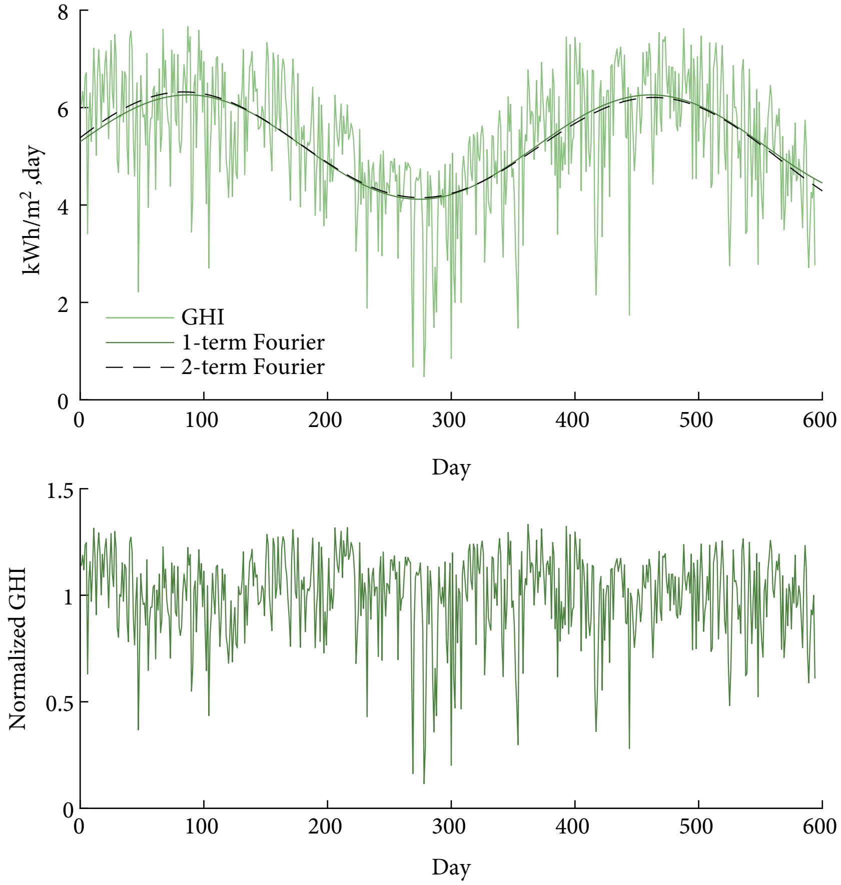

FIG. 3.2
Examples of Fourier series, in the upper graph, fitted to daily GHI data from the rotating shadow band radiometer in the Oahu Solar Measurement Grid between 17 March 2010 and 31 October 2011 (Sengupta and Andreas, 2020). The lower graph shows the daily GHI normalized by the 1-term Fourier series.

and modeling on different months, they applied specific transformations for each monthly period to obtain one full-year stationary time series. A similar approach with monthly transformations of clearness index data was also used by Graham *et al.* (1988). In later studies, Bright *et al.* (2015) used Markov chains representing different sub-annual periods to capture seasonal variations in climate parameters affecting solar irradiance.

Another related approach is to make this partitioning of data not in the temporal domain, but, observing that the statistical features of the clearness index are strongly related to the average over some period of time, sort the data into bins for the temporally averaged clearness index. Recalling that stationarity in the strict sense requires that the probability distribution of the data does not change over time, the reason for doing this is to separate those parts of the data that clearly have very different probability distributions.

For example, Aguiar and Collares-Pereira (1992) determined the probability distributions for the hourly clearness index for bins of the daily clearness index. This allowed applying transformations to the hourly time series of each day to make the full-year hourly data stationary. Graham and Hollands (1990) similarly studied the statistical properties of the hourly clearness index in bins of daily clearness index, but instead of making transformations of the clearness index data they determined explicit analytical functions describing how distribution parameters depend on the daily clearness index. The same approach has been used for higher-resolution binning in many later studies, including Widén *et al.* (2017). Grantham *et al.* (2017) also used binning of higher-resolution data (five-minute data within bins of hourly clear-sky index) in their bootstrapping approach. Also, recently, Frimane *et al.* (2019) developed mathematical classifications of daily clearness indices for modeling sub-daily features.

An example of binning of clear-sky index (CSI) data in this sense is shown in Fig. 3.3, where each hour of the measurement period was sorted into bins based on the average hourly CSI. Within each such hourly CSI bin, empirical probability distributions of the 15-second data within each hour were

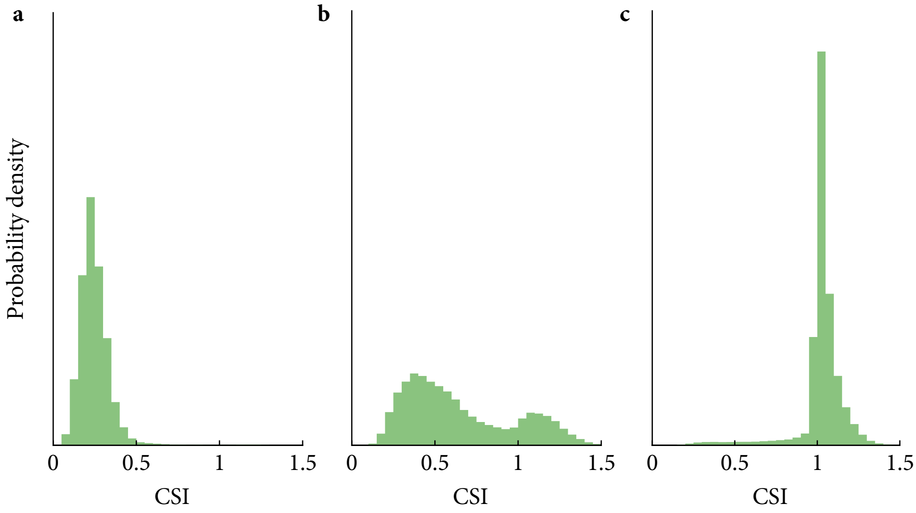

FIG. 3.3
Empirical probability distributions of the 15-second clear-sky index (CSI) in different bins of hourly average CSI. The hourly CSI is between 0.2 and 0.3 in (a), between 0.6 and 0.7 in (b), and between 1.0 and 1.1 in (c). Based on data from the Oahu solar measurement grid (Sengupta and Andreas, 2020).

then determined. As can be seen, the probability distributions are very different between the bins, not only in terms of mean value and variance, but also in the number of peaks in the probability density. In this case it would make sense to analyze the bins separately. This can be achieved, for example, by fitting a specific model of the 15-second CSI in each bin.

3.3.4 Probability distribution transformations

A transform is, in general, a mathematical function that can be used to format data to some more practical form. In the context of synthetic irradiance modeling, we look for transforms that make irradiance data easier to reproduce with a mathematical model. Transforming data so that they get a standard Gaussian distribution instead of a more complex probability distribution is a typical example.

Transforms can also be used to achieve stationarity. In the case of data separated into time intervals or bins, as described earlier, instead of fitting different models within each stationary interval or bin, the data in each partition could be transformed using partition-specific transforms so that the whole dataset ends up having the same distribution.

If the data are already normally distributed, or approximately normally distributed, with some mean value and standard deviation, calculating the z-score—that is, subtracting the mean and dividing by the standard deviation—will make sure the data in all partitions have the same distribution:

$$\hat{X} = \frac{X - \mu}{\sigma}. \tag{3.4}$$

For example, Aguiar and Collares-Pereira (1992) used this normalization method applied to the distributions of the hourly clearness index within each bin of daily clearness to make the whole dataset stationary.

For more complex distributions, a probability distribution transform may be applied. Any random variable X can be transformed to a uniformly distributed random variable U on the interval [0,1] by applying to it its own cumulative distribution function (CDF), that is,

$$U = F_X(X). \tag{3.5}$$

If this type of transformation is applied to two differently distributed random variables X and Y, both will yield identical uniformly distributed random variables, which can be equated, so that

$$F_X(X) = F_Y(Y), \tag{3.6}$$

and, consequently,

$$Y = F_Y^{-1}(F_X(X)). \tag{3.7}$$

The process of transforming X to Y is illustrated in Fig. 3.4, where a CSI time series is first transformed by its own empirical CDF (F_x) to become uniformly distributed, and then transformed through the inverse Gaussian CDF (F_Y^{-1}) to become normally distributed.

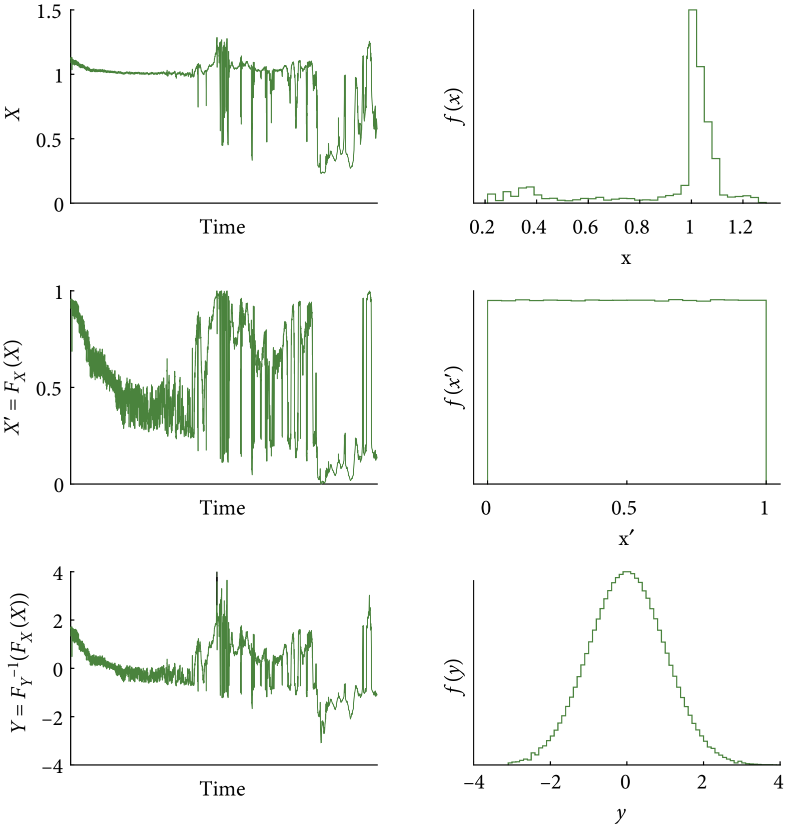

FIG. 3.4
Step-by-step probability distribution transformation, from top to bottom, of a CSI time series to make it normally distributed. The figures on the left show the original and transformed 1-second CSI time series, and the ones on the right show the corresponding empirical probability distributions of the data. In the first transformation step the empirical CDF was applied, and in the second step the inverse Gaussian CDF. The data are the same as in Fig. 3.1.

This transformation scheme was used by Graham *et al.* (1988) to transform daily clearness index data in monthly time intervals into data with a standard normal distribution, so that an autoregressive model could be fitted. In Graham and Hollands (1990), the hourly random component of the clearness index, assumed beta distributed, was transformed to a normally distributed random variable in the same way. Probability distribution transforms are also used in the copula method (see Sec. 3.7.2) and thus form part of many recent approaches to synthetic irradiance data generation. It should be noted that linear correlation is usually not preserved during these transformations (Widén *et al.*, 2017). Care should therefore be taken to evaluate autocorrelation in transformed time series.

3.4 PROBABILITY DISTRIBUTION MODELS

Probability distributions play an important role in mathematical modeling of solar irradiance. As seen in Sec. 3.3.4, probability distributions are used for transforming data to uniform or Gaussian distributions in order to utilize their specific properties. Also, as will be seen further on, probability distributions are used in, e.g., time series modeling, copula modeling, and random sampling. Recent studies on characterization of probability distributions for solar irradiance, many of which are cited in the following, have shown that solar irradiance can be consistently modeled by certain probability distributions that depend on parameters such as solar elevation and cloud cover, and not the geographic location per se. This suggests that solar irradiance should, in theory, be possible to fully characterize, model, and reproduce independently of location. Thus, the importance of probability distribution modeling to synthetic irradiance modeling cannot be understated.

Probability distribution estimation and modeling encompasses a variety of methods, where estimating probability distribution family parameters and estimating mixture distribution parameters have been particularly used in the synthetic solar irradiance generation literature. As solar radiation depends on the sun's position in the sky and on local climate and weather, the probability distributions may be dependent on geographic location as well as time of day and season. However, the distribution for the clear-sky index shows some general and more or less universal characteristics at different resolutions, such as bimodality or even trimodality at high resolutions. An important part of model development for synthetic generation of irradiance data is therefore to find the right distribution for the intended time resolution, fitting some set of training data.

In distribution fitting, commonly a distribution family is decided upon based on observation of a dataset to be modeled, and fitting a probability density function (PDF) means determination of parameters by some estimation technique (Casella and Berger, 2002). There are several methods that can be used, where maximum likelihood estimation (MLE) or the method of moments (MoM) are commonly used. MLE is a method for estimating the parameters of a probability distribution by maximizing the likelihood function (McLahlan and Peel, 2000, p. 40). MoM, on the other hand, estimates parameters of a predefined probability distribution function *f* by equating *N* statistical moments with different orders of the sampled values with the same moment orders of the distribution *f*, and solving for the parameters (Kotz and Johnson, 1985).

Beyond single-family distributions, the probability distribution may also consist of a combination of probability distributions, which is then called a mixture distribution. Formally, a mixture distribution $f(x)$ may be defined based on a combination of the probability distributions $f_1(x), f_2(x), \ldots, f_N(x)$ according to (McLahlan and Peel, 2000, p. 6)

$$f(x) = \sum_{i=1}^{N} \pi_i f_i(x), \tag{3.8}$$

where $0 \leq \pi_i \leq 1$ are weights of the function such that $\sum \pi_i = 1$ when summed over all *i*. A model with two distributions is called a bimodal distribution and a model with three distributions is called a trimodal distribution. The special case of using Gaussian probability distribution models is called a Gaussian mixture model (GMM) (Murphy, 2012, p. 339). Fitting distributions may be useful in certain specific model types where a distribution has to be set and synthetic data are generated by sampling from this distribution, one example being copula modeling.

For synthetic solar irradiance data generation, various probability distribution models of GHI, clearness index, and clear-sky index have been developed. Munkhammar and Widén (2018a) fitted normal, lognormal, and polynomial probability distributions to the clear-sky index in a two-state model. In that study MLE was used to fit normal and lognormal distributions, while the MoM was used to fit polynomial distributions. Fernandez-Peruchena and Bernadoz (2015) fitted Boltzmann statistics to GHI under conditioning of optical air mass by using the Levenberg–Marquardt algorithm to search for coefficients that minimize chi-square. Bright *et al.* (2015), Bright *et al.* (2017), and Smith *et al.* (2017) fitted various types of distribution by MLE to large samples of hourly clear-sky index data, resulting in generalized gamma and Burr type III distributions. The clear-sky index was partitioned into bins of the cloud cover fraction (0–9) and increments of solar elevation angle (0°, 10°, …, 90°), resulting in 81 fitted distributions covering all cloud conditions and solar angles. Zhang *et al.* (2018a) fitted a combination of normal and beta distributions and used these in a Cholesky decomposition model. Frimane *et al.* (2019) fitted a GMM to model the clearness index, and Shepero *et al.* (2019) fitted a GMM to model the clear-sky index using the Baum–Welch algorithm. It should be noted that the studies by Frimane *et al.* (2019) and Shepero *et al.* (2019) used different underlying correlation models. Widén *et al.* (2017) fitted bimodal and trimodal GMMs by MLE to the clear-sky index, and then used these as marginal distributions in a spatial copula model. The copula-based spatiotemporal model in Widén and Munkhammar (2019b) also utilized a bimodal GMM, also with the use of MLE.

Figure 3.5 shows examples of trimodal GMMs based on the model in Widén *et al.* (2017). This model was identified based on GMM fitting in average daily CSI bins, similar to the data in Fig. 3.3 but for bins of daily CSI instead of hourly CSI. In the nomenclature of Eq. (3.8), $N = 3$ and the probability distributions $f_1(x)$, $f_2(x)$, and $f_3(x)$ are Gaussian distributions, the mean values, and variances of which depend on the average daily CSI. These correspond to obscured sun ($f_1(x)$), unobscured sun in partly cloudy sky ($f_2(x)$), and unobscured sun in clear sky ($f_3(x)$). A higher daily CSI implies, naturally, a higher weight for the unobscured sun states. In addition, the weights π_1, π_2, and π_3 also depend on the daily CSI (note, though, that in Fig. 3.5(a) there are only two non-zero weights, as no clear sky is expected at this low daily CSI).

It should be noted that the preceding account mainly focused on cases where distributions have been modeled explicitly and connected to some synthetic data generator, not studies limited to only distribution fitting to data, or analysis of distributions as the outcome of a more complex model. Also, only probability distribution models for solar irradiance, clearness index, or clear-sky index were considered.

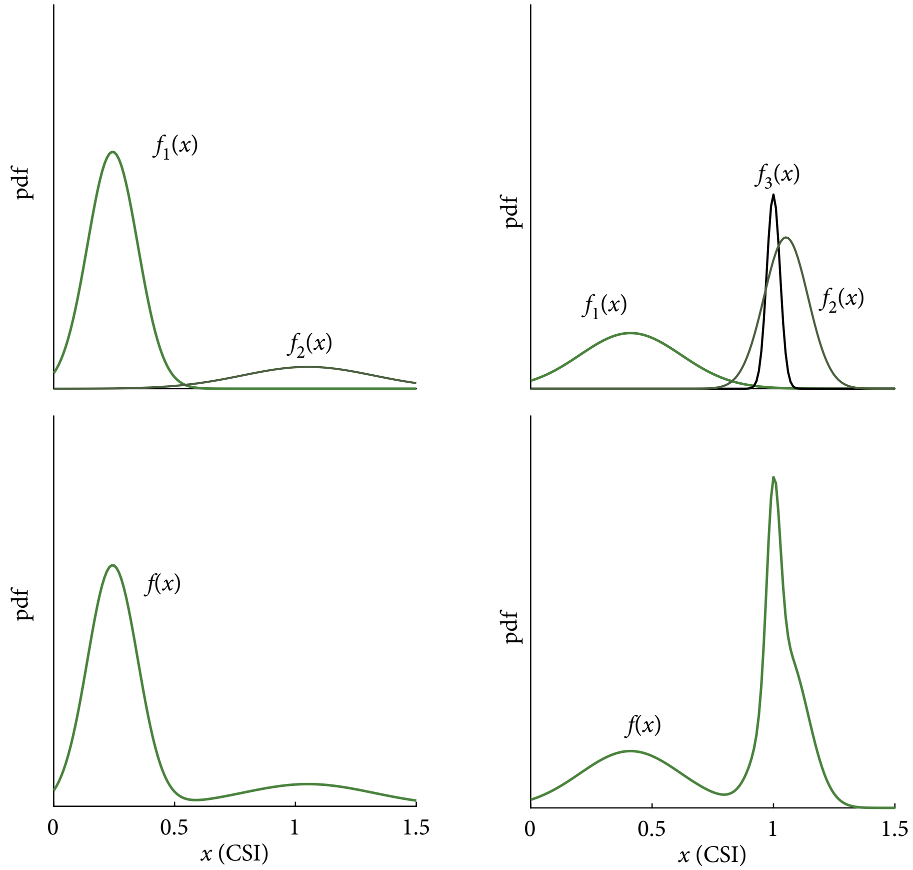

FIG. 3.5

Three-state Gaussian mixture distribution model of the instantaneous clear-sky index, based on the model in Widén *et al.* (2017). The upper figures show the individual Gaussian distributions and the lower ones the combined mixture distribution. The daily CSI is 0.4 in the figures to the left and 0.8 in the figures to the right.

3.5 AUTOREGRESSIVE AND RELATED PROCESSES

Several synthetic generators for solar irradiance time series have been based on autoregressive models, especially many of the early approaches. Also other, related models have been proposed as extensions or generalizations of the autoregressive process.

3.5.1 Autoregressive models

Autoregressive (AR) models are widely used for describing processes in both natural and human-made systems. They can be used for modeling time series where the values depend in a systematic way on previous values, yet also exhibit some degree of randomness. More specifically, an AR model is a random process where the value in each time step depends linearly on the previous values, and where randomness is added in each time step by independent white noise terms. An AR model of order p, denoted AR(p), is defined as (Shumway and Stoffer, 2011)

$$Y(t) = c + \sum_{i=1}^{p} \phi_i Y(t-i) + \epsilon(t), \tag{3.9}$$

where c is a constant, ϕ_i are model parameters, and $\epsilon(t)$ is random white noise at time t, i.e., uncorrelated random variables, generally assumed to be normally distributed with zero mean and some finite variance. An AR process thus depends on previous values in the time series (p time steps back), i.e., it has *memory*, combined with an element of randomness, depending on the distribution of the white noise component.

Identifying an AR model for a time series is usually done in three main steps, which most of the studies on AR-based synthetic irradiance generators adhere to: (1) transforming the irradiance data to stationary, normally distributed data; (2) determining the suitable order p of the AR model; and (3) fitting the model parameters to the data.

Methods for the first step were discussed in Sec. 3.3. For the second step, in many studies the partial autocorrelation function is usually investigated. The partial autocorrelation for a lag k is the autocorrelation with the dependence on lags shorter than k removed; hence the partial autocorrelation for an AR(p) model is zero for lags beyond k (Box *et al.*, 2016). The reasoning is that if the partial autocorrelation for a lag higher than τ is sufficiently close to zero, then τ can be considered an appropriate value for the order p. However, it should be noted that even though this approach is often used, a formally more correct method is to choose the model for which the so-called Akaike's information criterion (AIC) is minimized (Shumway and Stoffer, 2011).

Bertoli *et al.* (1981) and Amato *et al.* (1985), investigating autocorrelations for daily irradiance time series for Italy, concluded that an AR(1) process is sufficient for reproducing daily irradiance time series. The same conclusion was reached by Brinkworth (1977) for daily United Kingdom irradiance data. A much more thorough analysis of partial autocorrelation functions, and considering more alternative models (including ARMA; see Sec. 3.5.2), was undertaken by Graham *et al.* (1988) for Canadian irradiance data, but came to the same conclusion; time series of daily solar irradiance, if properly transformed for stationarity and Gaussianity, can be accurately modeled with an AR(1) process on the following form:

$$Y(t) = \phi Y(t-1) + \epsilon(t). \tag{3.10}$$

Synthetic time series generated with an AR model have a Gaussian distribution, but can be transformed back to CSI by applying the process that was shown in Fig. 3.4 in reverse. This is outlined in Fig. 3.6, where time series generated by two AR(1) models are first transformed to become uniformly distributed by applying the Gaussian CDF, and then transformed into CSI by applying the CSI inverse CDF, in this case based on the distribution in Fig. 3.5(b). The two AR(1) processes differ in the ϕ parameter, which is set to $\phi = 0.01$ and $\phi = 0.9$, respectively, in the two models. The higher the ϕ parameter, the higher the autocorrelation in the time series; as can be clearly seen, the time series with $\phi = 0.01$ is less rapidly fluctuating. For both time series, the variance of the Gaussian white noise term $\epsilon(t)$ was set to $\sigma^2 = 1 - \phi^2$, which makes the resulting time series $Y(t)$ normally distributed with zero mean and variance equal to one.

This type of AR(1) modeling, with probability distribution transforms similar to the ones outlined here, was applied to hourly solar irradiance data with satisfying results and formed the basis of both the Graham and Hollands (1990) model (the GH model in short) and the so-called time-dependent, autoregressive, Gaussian (TAG) model by Aguiar and Collares-Pereira (1992). These two models can still be considered absolute state of the art in synthetic generation of hourly data and are of practical importance, as they are implemented in widely used solar design and simulation software.[1]

3.5.2 Other time series models

Although the AR(1) process has been convincingly shown to accurately reproduce observed variability in hourly solar irradiance, other similar models have also been proposed for synthetic hourly time series, based on the generalized ARMA process, which, in addition to an autoregressive part as in the expression above (Eq. 3.9), also contains a moving-average part:

$$X(t) = c + \sum_{i=1}^{p} \phi_i X(t-i) + \epsilon(t) + \sum_{i=1}^{q} \theta_i \epsilon(t-i), \tag{3.11}$$

where θ_i are model parameters defining a moving average over the random white noise terms. This moving average term introduces a linear dependence not only between previous process values but directly between the stochastic terms. This makes the ARMA process behave differently than the AR process, for example, in terms of autocorrelation and partial autocorrelation. It may therefore be that an ARMA process, an AR process, or even a pure MA process, best describes the time series at hand.

[1] At the time of writing, the GH model is used for generation of synthetic hourly solar irradiance data from monthly averages in the microgrid simulation software HOMER Pro and the TAG model is used for generation of hourly irradiance data from daily values in the photovoltaic system design software PVsyst.

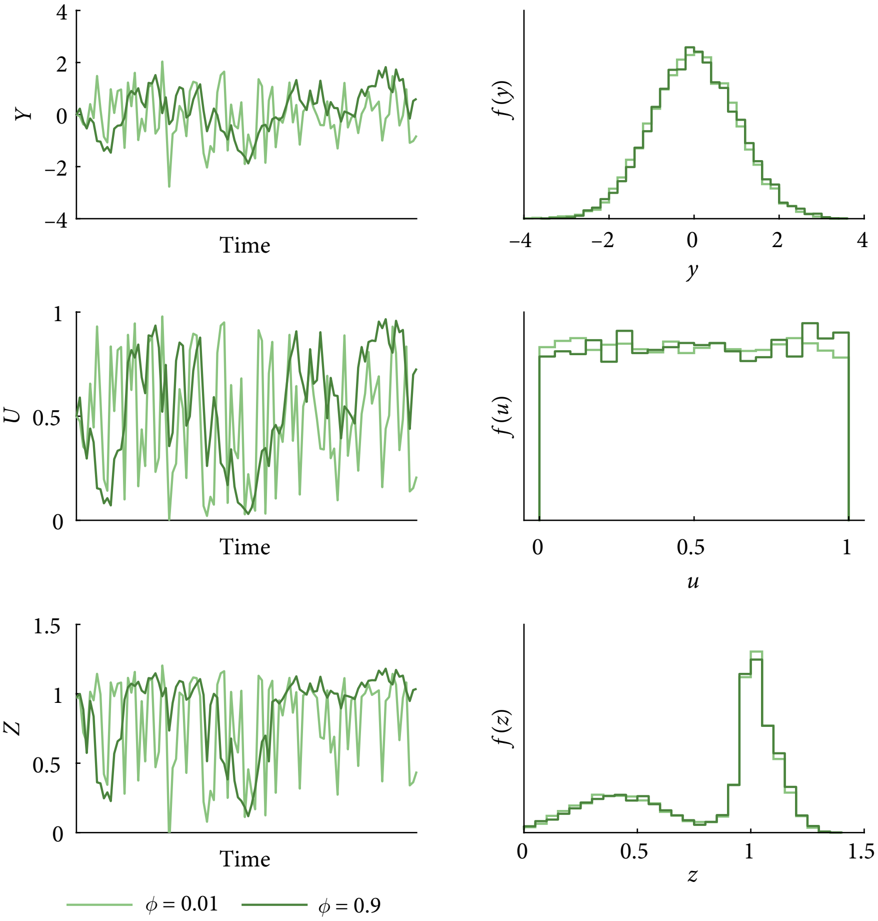

FIG. 3.6
Example of two synthetic time series based on the AR(1) process, transformed stepwise from standard Gaussian to CSI distributions using probability distribution transforms. The figures show, from top to bottom, the original processes, the processes transformed by the standard Gaussian CDF, and the processes transformed again by the inverse CDF of the CSI, based on the CSI distribution in Fig. 3.5(b).

Vergara-Dominguez *et al.* (1985) used a method proposed by Jacobs and Lewis (1978a; 1978b) to generate random time series with a certain autocorrelation and marginal distribution, based on an ARMA process. Mora-Lopez and Sidrach-de-Cardoba (1998) used a seasonal (multiplicative) ARMA model, in which a first-order autoregressive process described the regular part and a first-order moving-average process described the seasonal part.

Another form of generalization of the AR process is the nonlinear autoregressive (NLAR) model, in which the process at time t depends on a number of previous process values and a random white noise term, but according to a nonlinear function:

$$X(t) = f(X(t-1),\ldots,X(t-p),\epsilon(t)). \tag{3.12}$$

For higher-resolution time series, Glasbey (2001) developed a NLAR model able to reproduce the bimodal marginal distribution of 30-s solar irradiance data.

3.6 MARKOV CHAIN MODELS

A Markov chain is a discrete-time stochastic process in which one out of a limited number of states is occupied at each time step and in which there exist probabilities for transition between states from one time step to the next. In solar irradiance modeling, these states could be, for example, irradiance levels or cloud amount.

Markov chain models can be divided into discrete-time and continuous-time Markov chains (Cinlar, 1975). A discrete-time Markov chain $S(t)$ is a discrete stochastic process X_t that, in each time step, occupies one state s_i in a number of defined states $s_1, s_2, \ldots, s_N$. It also satisfies the Markov property, which states that the process is only dependent on the previous time step and thereby it lacks memory. If we let the time steps be defined as $t = 1,\ldots,T$, and $i = 1,\ldots,N$ denotes the index determining which state the process is in at a given time step, then the probability that the process occupies a particular state s_i at time step t is (Cinlar, 1975)

$$p_i(t) = \text{Prob}(X_t = s_i). \tag{3.13}$$

Since the Markov chain is based on a particular set of states we have that

$$\sum_{i=1}^{N} p_i(t) = 1. \tag{3.14}$$

The transitions from one state to another are defined by a transition matrix:

$$P_{ij} \equiv \text{Prob}(X_{t+1} = s_i | X_t = s_j), \tag{3.15}$$

where $i = 1,\ldots,N$ and $j = 1,\ldots,N$, and therefore P is a square matrix.

A Markov chain that is based on a time-independent transition matrix, like in Eq. (3.15), is called homogenous, and a Markov chain based on a time-dependent transition matrix is called inhomogenous. The inhomogenous case then introduces a time-dependence in the transition matrix such that $P_{ij} = P_{ij}(t)$. A Markov chain for which the conditional distribution of a variable only depends on one previous observation is called a first-order Markov chain, and a Markov chain that depends on M previous observations is called an Mth order Markov chain (Bishop, 2006, p. 609).

As an extension of a traditional Markov-chain-based model, a hidden Markov model (HMM) is a Markov chain with states that are not directly visible in the model output. In fact, we have already encountered the concept of hidden states in the discussion on CSI mixture models. These have a finite number of states, corresponding to sky states such as obscured or unobscured sun, but within each state there is a continuous range of CSI values that the model can assume. These ranges also overlap between the states, so that a CSI value does not unanimously correspond to a specific state. Here, the underlying sky states are hidden, while the CSI is visible.

Indeed, a HMM can be seen as a generalization of a mixture distribution (MacDonald and Zucchini, 1997; and Bishop, 2006), where a Markov chain is used for selection of mixture component. Formally, given an N component mixture distribution with probability densities $p_1(x), p_2(x), \ldots, p_N(x)$, then this model can be equipped with an N-state Markov chain with states $s_1, s_2, \ldots, s_N$. In this construction each state is associated with a mixture component in such a way that the choice of mixture component is not selected independently, as it would be in a regular mixture distribution model approach, but instead it depends on the Markov-chain-based choice of components from the previous observation (Bishop, 2006). A two-state example would be a Markov chain with two states, a 2 × 2 transition matrix combined with a mixture distribution $f(x)$ consisting of a mixture of distributions $f_1(x)$ and $f_2(x)$:

$$f(x) = \pi_1 f_1(x) + \pi_2 f_2(x), \tag{3.16}$$

where the weights π_1 and π_2 represent the stationary distribution of the Markov chain. The stationary distribution π_i for a Markov chain satisfies

$$\pi_i P_{ij} = \pi_i, \tag{3.17}$$

which implies that the stationary distribution π_i is invariant when multiplied with a transition matrix. For more detailed information on HMMs, see MacDonald and Zucchini (1997), Bishop (2006), and Murphy (2012).

Markov chains have been used extensively in the solar irradiance modeling literature. Solar irradiance and its normalized versions of clearness index and clear-sky index have been modeled in both temporal (Aguiar *et al.*, 1988; Morf, 1998; Poggi *et al.*, 2000; Wegener *et al.*, 2012; Morf, 2013; Ngoko *et al.*, 2014; Bright *et al.*, 2015; and Munkhammar and Widén, 2018a, 2018b, Bright, 2019) and spatiotemporal (Bright *et al.*, 2017; and Munkhammar and Widén, 2019) settings.

The clearness index was modeled by Aguiar *et al.* (1988) for various climate regions and for data resolution ranging from hourly to one month, combining a Markov chain with analytical expressions of the clearness index. The two-state Markov model STSIM consisting of bright sunshine and cloudy conditions was developed by Morf (1998). Poggi *et al.* (2000) modeled the clearness index using a 20-state Markov chain model combined with a shifted negative binomial distribution model; the model was then used to make an hour-resolution clearness index generator. An HMM was developed by Wegener *et al.* (2012), where it was combined with a wavelet decomposition for the purpose of downscaling data. Morf (2013) combined Markov chains with a stochastic insolation function (SIF) to generate horizontal synthetic solar irradiance data. First- and second-order Markov chains for generating synthetic clearness index data were developed by Ngoko *et al.* (2014) and applied to one-minute resolution solar irradiance data. An extensive Markov chain model including weather observation data and cloud modeling was developed by Bright *et al.* (2015; 2017).

HMMs, in this case Markov chains combined with mutually disjoint mixture distributions, for generating synthetic clear-sky index data were developed for a two-state model by Munkhammar

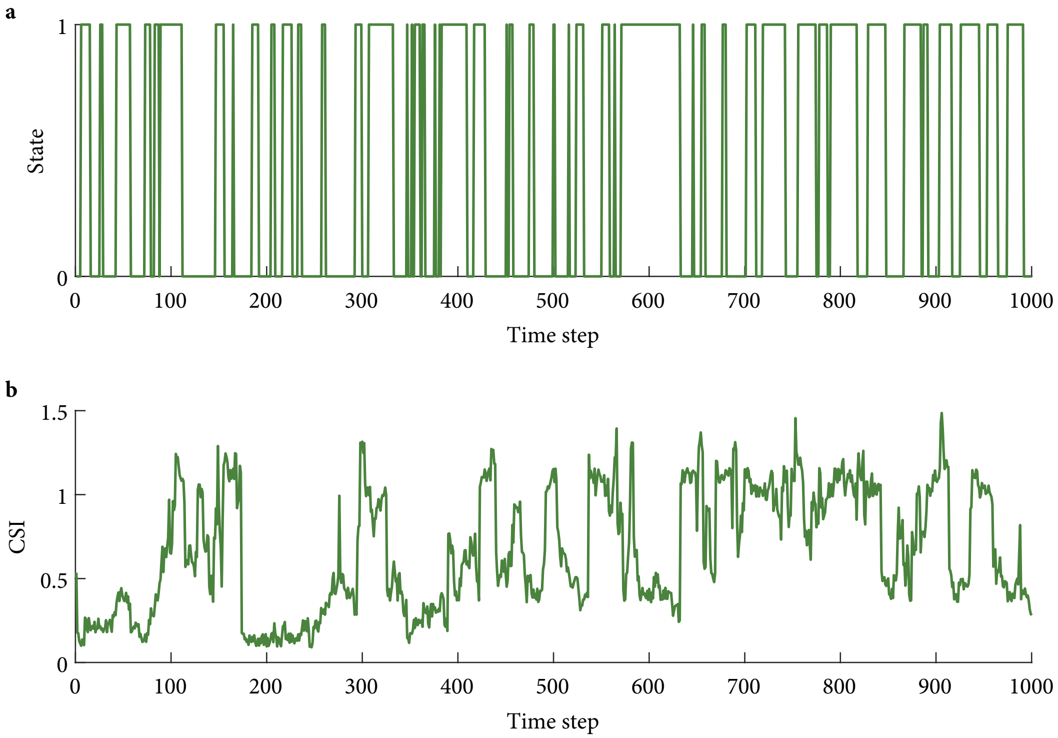

FIG. 3.7
Example of time series generated by two Markov chain models. In (a) a simple two-state model of cloud cover (0 = cloudy, 1 = clear), in (b) the *N*-state model (Munkhammar and Widén, 2018b) with *N* = 20.

and Widén (2018a; 2018b), who developed an N-state model as well. Both studies utilized minute resolution data for two different climatic conditions. In Shepero *et al.* (2019) a Gaussian mixture distribution HMM was developed for three states with connections to analytical expressions and also generally for N states. The model was applied to minute resolution data.

Of the few generators of spatiotemporal irradiance data, Bright *et al.* (2017) combined Markov chains with a cloud sampling algorithm and connected this to meteorological variables in order to produce synthetic spatiotemporal solar irradiance data. Munkhammar and Widén (2019) developed a multidimensional Markov chain model for generating synthetic clear-sky index.

Two examples of realizations of CSI Markov chains are shown in Fig. 3.7. The upper figure shows an example output from a simple two-state chain with a 90% probability of staying in the present state. This model could be used for modeling binary sky states such as obscured versus unobscured sun. As mentioned earlier, this is the model type that was used in Morf (1998). The lower figure shows an example output from the N-state model of Munkhammar and Widén (2018b) with $N = 20$. In this case, the CSI range is divided into 20 equally wide bins, each such bin corresponding to a state in the Markov chain. The chain transitions between these states with specific transition probabilities, and within each state CSI values are drawn from uniform distributions over the bin range.

3.7 MULTIVARIATE DISTRIBUTIONS AND COPULAS

An alternative to models that sequentially generate new synthetic time series values based on one or more previous values, as in the case of both AR processes and Markov chains, is to create models where each time step in a time series, or points in both space and time, are treated as individual dimensions in a multivariate distribution from which correlated samples are drawn. This section briefly outlines this concept and how to synthetically generate data from such a model.

3.7.1 Multivariate distributions

An N-dimensional multivariate distribution can be defined as a random vector of univariate stochastic variables $X = \{X_1, X_2, \ldots, X_N\}$ with the joint CDF (Mardia *et al.*, 1979, p. 26):

$$F(x_1, x_2, \ldots, x_N) = P(X_1 < x_1, \ldots, X_N < x_N), \tag{3.18}$$

of the real variables $x_1, x_2, \ldots, x_N$. The CDF F can be defined as an integral of a probability density distribution f over all variables (Mardia *et al.*, 1979, p. 26) or the probability density distribution f can be expressed as a derivative of F over all variables (Hazewinkel, 2001):

$$f(x_1, x_2, \ldots, x_N) = \frac{\partial^N F(x_1, x_2, \ldots, x_N)}{\partial x_1 \partial x_2 \ldots \partial x_N}. \tag{3.19}$$

The distribution of any of these random variables X_i relative to the multivariate distribution is called a marginal distribution. The marginal distributions are completely defined by the multivariate distribution, and when $X_1, X_2, \ldots, X_N$ are independent, the multivariate CDF becomes (Hazewinkel, 2001)

$$F(x_1, x_2, \ldots, x_N) = F_1(x_1)F_2(x_2) \cdot \ldots \cdot F_N(x_N), \tag{3.20}$$

where each F_i is a marginal distribution of F. In solar irradiance modeling, multidimensional probability distributions typically represent irradiance—or some normalized version—over a set of time steps, over a set of points in space, or points in space and time.

An example of a multivariate distribution of high relevance for synthetic irradiance modeling is the multivariate Gaussian distribution. This distribution over the variables X_1, X_2, …, X_N is characterized by the mean values $\mu = (\mu_1, \ldots, \mu_N)^T$ of the variables along with an $N \times N$ covariance matrix $\sum$, with the variances of the variables on the diagonal and the off-diagonal elements being the covariances (Miller, 1964, p. 20). The bivariate case, i.e., for $N = 2$, is shown in Fig. 3.8 for different correlations between the variables. When the correlation is zero, X_1 and X_2 are independent and can be treated as two separate univariate Gaussian variables. When the correlation is positive or negative, the distribution is concentrated around the lines $X_1 = X_2$ and $X_1 = -X_2$, respectively.

Samples can be drawn from a multivariate Gaussian distribution using Cholesky decomposition, in which the covariance matrix is decomposed into the product of a lower triangular matrix and its transpose (Golub and Ortega, 1992):

$$\sum = \mathbf{A}\mathbf{A}^T. \tag{3.21}$$

Given N independent samples $\mathbf{y} = (y_1, \ldots, y_N)^T$ from a standard univariate Gaussian distribution, the samples from the multivariate Gaussian can then be obtained as (Golub and Ortega, 1992)

$$\mathbf{x} = \mu + \mathbf{A}\mathbf{y}. \tag{3.22}$$

A recipe for generating synthetic CSI data based on the multivariate Gaussian distribution would therefore be first to define the N variables in the multivariate distribution (they could be points in time or in space, or both), second to identify covariances between these variables from CSI data transformed to normally distributed data using the transformation methods described in previous chapters, and then third to draw correlated samples from the distribution using the Cholesky decomposition approach outlined earlier. The samples could then be transformed back from Gaussian to CSI by once again applying probability distribution transforms; this last step is a special case of sampling from so-called copulas, which are described in the next section.

3.7.2 Copulas

For modeling a correlated multivariate probability distribution, the copula method can be useful. A copula is a multivariate cumulative distribution function for which the marginal probability distribution is uniform. Copulas can be used to model the dependence between stochastic variables with

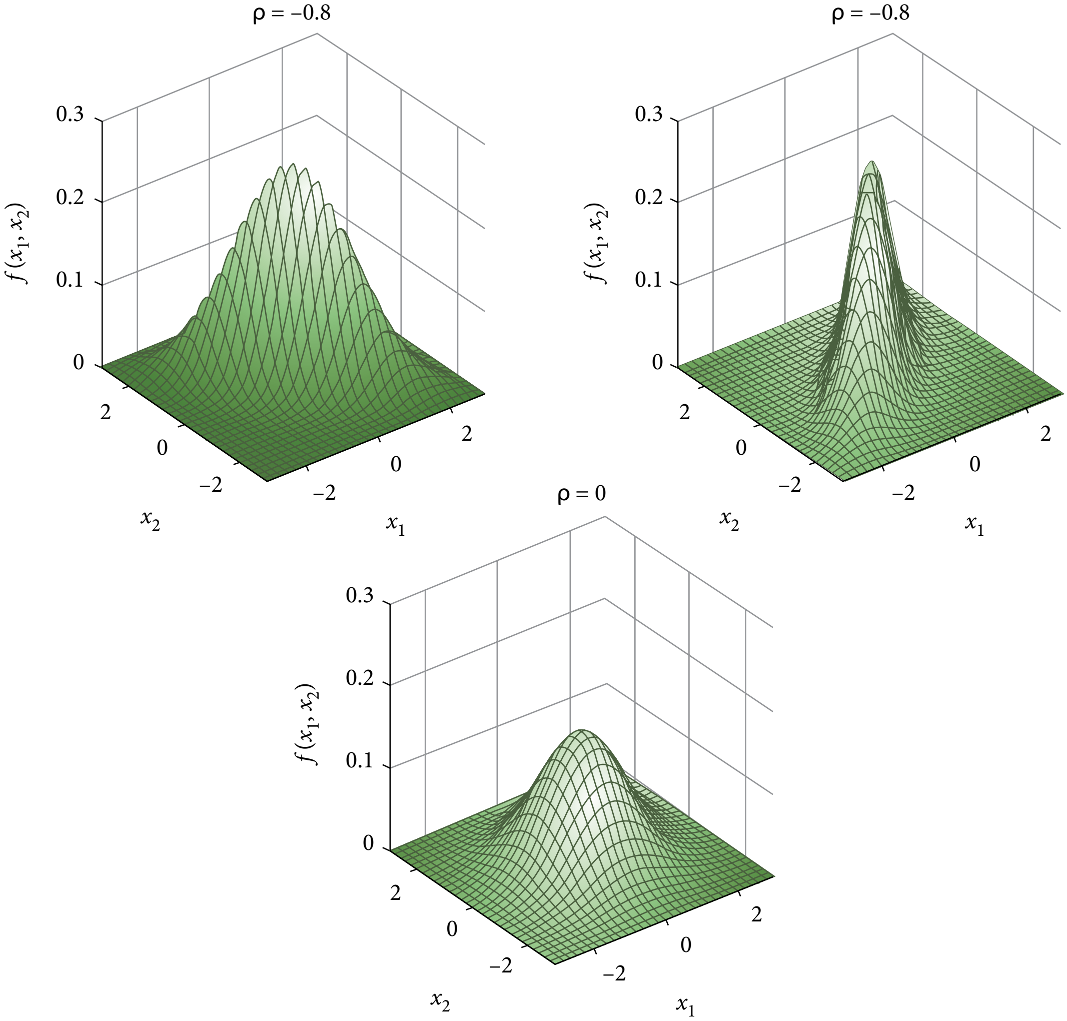

FIG. 3.8

Bivariate Gaussian distribution with different correlations between the two variables X_1 and X_2. Both means are zero and both variances are equal to one.

any arbitrary distribution. In particular, for multivariate distributions with complicated marginals, such as the multimodal distributions of high-resolution clear-sky index data, it is usually more feasible to connect the exact marginals through a copula with an appropriate correlation structure than trying to fit some specific multivariate distribution family to the data.

Formally, given N stochastic variables $X_1, X_2, \ldots, X_N$ with CDFs F_{X_1}, $F_{X_2}, \ldots, F_{X_N}$, according to Sklar's theorem, the existence of a copula is certified (Nelsen, 2006). That implies that the distributions $F_{X_1}, F_{X_2}, \ldots, F_{X_N}$ can be joined via a copula C if the copula can be expressed as

$$F_{X_1,X_2,\ldots,X_N}(X_1, X_2, \ldots, X_N)$$
$$= C(F_{X_1}(x_1), F_{X_2}(x_2), \ldots, F_{X_N}(x_N)). \qquad (3.23)$$

The copula function can be written according to

$$C(u_1, u_2, \ldots, u_N)$$
$$= F_{X_1,X_2,\ldots,X_N}(F_{X_1}^{-1}(u_1), F_{X_2}^{-1}(u_2), \ldots, F_{X_N}^{-1}(u_N)), \quad (3.24)$$

for $F_{X_i} = u_i$, $i = 1, 2, \ldots, N$, where $u_1, u_2, \ldots, u_N$ are realizations of uniform variables $U_1, U_2, \ldots, U_N$, respectively. The special case with two stochastic variables is called a bivariate copula.

An often-used copula is the Gaussian copula, which is defined based on the multivariate Gaussian distribution, discussed in Sec. 3.7.1. In this case, the CDFs $F_{X_1,X_2,\ldots,X_N}$ and $F_{X_i}^{-1}$ in Eq. (3.24) are the joint CDF of a multivariate Gaussian distribution with zero mean vector and a specified correlation matrix, and the inverse CDF of a standard univariate Gaussian distribution, respectively. This means that in order to draw samples from a Gaussian copula, one first draws an N-dimensional correlated sample from a multivariate Gaussian distribution as described in Sec. 3.7.1, and then applies the inverse CDF of the standard Gaussian distribution to these samples to make them uniformly distributed.

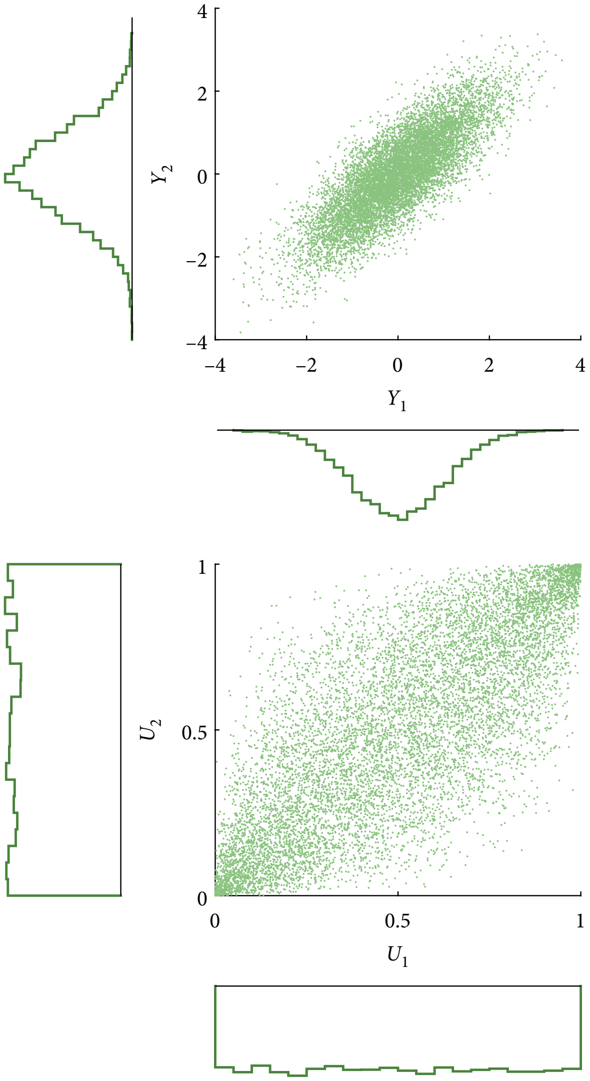

FIG. 3.9

Ten thousand samples from a two-dimensional Gaussian copula. The upper figure shows samples from a bivariate Gaussian distribution with zero means and correlation 0.8. The lower figure shows these samples transformed to uniformly distributed values through the inverse standard Gaussian CDF. The Gaussian and uniform marginals of the distributions are shown to the left of and below the scatterplots.

This is illustrated in Fig. 3.9, which shows scatterplots of 10 000 samples from a bivariate Gaussian copula. Note that each sample here only consists of two variables (for example, two time steps), whereas a useful CSI copula model would have many more dimensions; for example, for generating CSI time series between sunrise and sunset with a one-minute resolution, N would have to be in the order of 700–800.

Copula modeling of solar irradiance has focused on modeling the variability of the clear-sky index temporally (Munkhammar and Widén, 2017a; 2017b), spatially (Munkhammar and Widén, 2016; Widén *et al.*, 2017; and Munkhammar and Widén, 2017a) and spatiotemporally (Widén and Munkhammar, 2019b). The temporal model by Munkhammar and Widén (2017a; 2017b) utilized N time steps and an autocorrelation matrix for the clear-sky index and a Gaussian copula to generate N time step correlated time series. The input marginal distributions were modeled as empirical distributions of the clear-sky index for the datasets used. The models were evaluated for the different climatic conditions of Sweden and Hawaii, USA, for minute resolution. The spatial modeling (Munkhammar and Widén, 2016; Munkhammar and Widén, 2017a; and Widén *et al.*, 2017) was based on the correlation in the clear-sky index between N locations, combined with a Gaussian copula and empirical marginal distributions (Munkhammar and Widén, 2016; and Munkhammar and Widén, 2017a) and modeled distributions by Widén *et al.* (2017) to generate synthetic clear-sky index data for N locations. This was evaluated on minute resolution for different climatic regions of Norrköping, Sweden, and Hawaii, USA (Munkhammar and Widén, 2016; Munkhammar and Widén, 2017a; and Widén *et al.*, 2017).

A spatiotemporal model of the clear-sky index using copula was introduced in Widén and Munkhammar (2019b), where the virtual network approximation was used in combination with temporal autocorrelation and spatial correlation estimates to generate any number of outputs in space or time for a geographical region. This was evaluated using 15-s resolution irradiance data for 17 locations in the Oahu Solar Measurement Grid (Sengupta and Andreas, 2020).

3.8 MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE METHODS

Machine learning is concerned with developing methods for automatic pattern recognition and using this to predict data (Murphy, 2012, p. xxxvi). This section collects some of the methods involved so far in synthetic solar irradiance modeling that do not naturally fit into the other sections and are often counted as belonging to the fields of machine learning and artificial intelligence (AI).

3.8.1 Artificial neural networks

An artificial neural network (ANN) is based on a collection of connected units, defined as artificial neurons, which are connected in a network. A key issue is training an ANN, which in practice implies setting weight values associated with the connections (Murphy, 2012). Neural networks have been used in solar irradiance modeling (Hontoria *et al*., 2001). In terms of time-series prediction, the model relies on the assumption that there exists a function *f* that relates the series s_i over steps *i*:

$$s_{i+1} = f(s_{i-p+1},\ldots,s_n) \tag{3.25}$$

up to an index *N* for which $p < i < N$, and that *f* can be approximated by an multilayer perceptron (MLP). For more detailed information on neural networks and MLP, see Bishop (2006) and Murphy (2012, p. 563).

A few synthetic solar irradiance generators have used ANNs. In Hontoria *et al.* (2001) a feedforward–feedback neural network architecture was developed and used to generate hourly solar irradiance data series by modeling the clearness index. The study used hour resolution data from three different climatic regions in Spain to evaluate the model. In Mora-Lopez and Sidrach-de-Cardoba (2003), a probabilistic finite automata (PFA) model was developed to simulate hourly GHI.

3.8.2 Clustering

Clustering, as part of unsupervised learning, is the task of grouping together a set of objects so that the objects belong to a specified group and are more similar to each other than to those in other groups. Clustering was used in the Dirichlet process Gaussian mixture model (DPGMM) (Frimane *et al*., 2019), which is a non-parametric Bayesian (NPD) model with infinite parameter space. This model utilizes a mixture distribution model with Gaussian distributions combined with a Dirichlet process (Görür and Rasmussen, 2010). A key benefit of the NPD approach is an automatic adaption to correct complexity level and model size, which is useful when adapting to, for example, different climate conditions (Frimane *et al*., 2019). For a complete mathematical model outline and evaluation of DPGMMs, see Görür and Rasmussen (2010) and Frimane *et al.* (2019).

Clustering was also used in Peruchena *et al.* (2018) in order to generate high-frequency synthetic GHI time series coupled with DNI time series. A one-minute resolution was modeled and one-minute resolution data from Carpentras, France, was used to validate the model.

3.8.3 Bootstrapping

The bootstrap method creates datasets based on an existing dataset. If an original dataset consists of *N* data points $X = \{x_1, x_2, \ldots, x_N\}$, then it is possible to create a new dataset *Y* by drawing *N* random samples from *X* (with replacement), so that points in *X* may be replicated in *Y*, whereas other points in *X* may be absent from *Y* (Bishop, 2006, p. 23). This process may then be repeated *M* number of times to obtain *M* datasets with *N* data points in each. The general purpose of this process is to evaluate the statistical accuracy of parameter estimates when utilizing the variability of predictions between the different bootstrap datasets.

Bootstrapping was used for synthetic solar irradiance generation by Grantham *et al.* (2017), where five-minute resolution GHI and DNI clear-sky index values were generated from categories of hour resolution data, as a synthetic irradiance generator, by bootstrapping. One-minute resolution data from a set of locations in Australia was used to validate the model. Another bootstrapping approach was presented by Grantham *et al.* (2018), where synthetic daily and hourly GHI data were generated using non-parametric bootstrapping combined with Fourier series.

3.9 SUMMARY

This chapter has given an overview of established mathematical approaches for generating synthetic solar irradiance data, based on the most important scientific studies from the last decades.

The applicability of the models that were presented, which represent the state of the art in the field, should be emphasized. They are useful for a multitude of applications ranging from providing solar resource data for photovoltaic or solar thermal system sizing to modeling and designing future smart cities. The models need to be adopted and adapted for particular purposes, as well as thoroughly calibrated and validated for the location and application under consideration, of which more can be read in Chap. 4.

It should also be emphasized that these methods are not exhaustive and the field is developing rapidly, with particularly interesting development during the last decade. The most unexplored frontier is currently spatiotemporal modeling, which has merely been initiated. An outlook toward further development of the field is provided in Chap. 6.

ACKNOWLEDGMENTS

The authors wish to thank Dr. Mahmoud Shepero for comments on the draft. This project was partially funded by the project “Development and evaluation of forecasting models for solar power and electricity use over space and time,” funded by the Swedish Energy Agency (Grant No. 43055-1). This work also forms part of the Swedish strategic research program StandUp for Energy.

REFERENCES

Aguiar, R. and Collares-Pereira, M., “TAG: A time-dependent, autoregressive, Gaussian model for generating synthetic hourly radiation,” Solar Energy **49**, 167–174 (1992).

Aguiar, R. J., Collares-Pereira, M., and Conde, J. P., “Simple procedure for generating sequences of daily radiation values using a library of Markov transition matrices,” Solar Energy **40**, 269–279 (1988).

Amato, U., Andretta, A., Bartoli, B., Coluzzi, C., Coumo, V., and Serio, C., “Stochastic modelling of solar-radiation data,” Il Nuovo Cimento **8**, 248–258 (1985).

Balouktsis, A. and Tsalides, P. H., “Stochastic simulation model of hourly total solar radiation,” Solar Energy **37**, 119–126 (1986).

Bertoli, B., Coluzzi, B., Cuomo, V., Francesca, M., and Cerio, C., “Autocorrelation of daily global solar radiation,” Il Nuovo Cimento **C2**, 113–122 (1981).

Bishop, C. M., *Pattern Recognition and Machine Learning* (Springer, New York, 2006).

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. *Time Series Analysis: Forecasting and Control*, 5th ed. (Wiley, Hoboken, New Jersey, 2016).

Brabec, M., Badescu, V., and Paulescu, M., “Cloud shade by dynamic logistic modeling,” J. Appl. Stat. **41**, 1174–188 (2013).

Bright, J. M., “The impact of globally diverse GHI training data: Evaluation through application of a simple Markov chain downscaling methodology,” J. Renew. Sustain. Energy **11**, 1–21 (2019).

Bright, J. M., Smith, C. J., Taylor, P. G., and Crook, R., “Stochastic generation of synthetic minutely irradiance time series derived from mean hourly weather observation data,” Solar Energy **115**, 229–242 (2015).

Bright, J. M., Babacan, O., Kleissl, J., Taylor, P. G., and Crook, R., “A synthetic, spatially decorrelating solar irradiance generator and application to a LV grid model with high PV penetration,” Solar Energy **147**, 83–98 (2017).

Brinkworth, B. J., “Autocorrelation and stochastic modelling of insolation sequences,” Solar Energy **19**, 343–347 (1977).

Cai, C. and Aliprantis, D. C., “Cumulus cloud shadow model for analysis of power systems with photovoltaics,” IEEE Trans. Power Syst. **28**, 4496–4503 (2013).

“CAMS McClear Service for estimating irradiation under clear-sky,” http://www.soda-pro.com/web-services/radiation/cams-mcclear (accessed 29 August 2020).

Casella, G. and Berger, R. L., *Statistical Inference*, 2nd ed. (Duxbury Press, Belmont, 2002).

Cinlar, E., *Introduction to Stochastic Processes* (Prentice-Hall, New Jersey, 1975).

Exell, R. H. B., “The fluctuation of solar radiation in Thailand,” Solar Energy **18**, 549–554 (1976).

Exell, R. H. B., “A mathematical model for solar radiation in South-East Asia (Thailand),” Solar Energy **26**, 161–168 (1981).

Fernandez-Peruchena, C. M. and Bernadoz, A., “A comparison of one-minute probability density distributions of global horizontal solar irradiance conditioned to the optical airmass and hourly averages in different climate zones,” Solar Energy **112**, 425–436 (2015).

Frimane, A., Soubdhan, T., Bright, J. M., and Aggour, M., “Nonparametric Bayesian-based recognition of solar irradiance conditions: Application to the generation of high temporal resolution synthetic solar irradiance data,” Solar Energy **182**, 462–479 (2019).

Glasbey, C. A., “Non-linear autoregressive time series with multivariate Gaussians as marginal distributions,” Appl. Stat. **50**, 143–154 (2001).

Golub, G. H. and Ortega, J. M., *Scientific Computing and Differential Equations: An Introduction to Numerical Methods* (Academic Press, San Diego, CA,1992).

Görür, D. and Rasmussen, C. E., “Dirichlet process Gaussian mixture models: Choice of the base distribution,” J. Comput. Sci. Technol. **25**, 653–664 (2010).

Graham, V. A. and Hollands, K. G. T., “A method to generate synthetic hourly solar radiation globally,” Solar Energy **44**, 331–341 (1990).

Graham, V. A., Hollands, K. G. T., and Unny, T. E., “A time series model for Kt with application to global synthetic weather generation,” Solar Energy **40**, 83–92 (1988).

Grantham, A. P., Pudney, P. J., Ward, L. A., Belusko, M., and Boland, J. W., “Generating synthetic five-minute solar irradiance values from hourly observations,” Solar Energy **147**, 209–221 (2017).

Grantham, A. P., Pudney, P. J., and Boland, J. W., “Generating synthetic sequences of global horizontal irradiation,” Solar Energy **162**, 500–509 (2018).

Hazewinkel, M. (ed.), “Joint distribution,” in *Encyclopedia of Mathematics* (Springer, New York, 2001).

Hontoria, L., Aguilera, J., Riesco, J., and Zufiria, P., “Recurrent neural supervised models for generating solar radiation synthetic series,” J. Robotic Intell. Syst. **31**, 201–221 (2001).

Jacobs, P. A. and Lewis, P. A., “Discrete time series generated by mixtures. I: Correlational and runs properties,” J. Roy. Stat. Soc. **B40**, 94–97 (1978a) .

Jacobs, P. A. and Lewis, P. A., “Discrete time series generated by mixtures. II: Asymptotic properties,” J. Roy. Stat. Soc. **B40**, 222–228 (1978b).

Jazayeri, M., Jazayeri, K., and Uysal, S., “Generation of spatially dispersed irradiance time-series based on real cloud patterns,” Solar Energy **158**, 977–994 (2017).

Kotz, S. and Johnson, N. L., *Encyclopedia of Statistical Sciences*, Volume 5 (John Wiley & Sons, 1985).

MacDonald, I. L. and Zucchini, W., *Hidden Markov and Other Models for Discrete-Valued Time Series, Monographs on Statistics and Applied Probability 70* (St. Edmundsbury Press, Bury St Edmunds, Suffolk, Great Britain, 1997).

Mardia, K. V., Kent, J. T., and Bibby, J. M., *Multivariate Analysis* (Academic Press, London, 1979).

McLahlan, G. and Peel, D., *Finite Mixture Models,* Wiley Series in Probability and Statistics (John Wiley & Sons, 2000).

Mellit, A., Benghanem, M., Hadj Arab, A., and Guessoum, A., “A simplified model for generating sequences of global solar radiation data for isolated sites: Using artificial neural network and a library of Markov transition matrices approach,” Solar Energy **79**, 469–482 (2005).

Miller, K. S., *Multidimensional Gaussian Distributions,* SIAM Series in Applied Mathematics (John Wiley and Sons, 1964).

Mohandes, M., Rehman, S., and Halawani, T. O., “Estimation of global solar radiation using artificial neural networks,” Renew. Energy **14**, 179–184 (1998).

Mora-Lopez, L. and Sidrach-de-Cardoba, B., “Multiplicative ARMA models to generate hourly series of global irradiation,” Solar Energy **63**, 283–291 (1998).

Mora-Lopez, L. and Sidrach-de-Cardoba, M., “Using probabilistic finite automata to simulate hourly series of global radiation,” Solar Energy **74**, 235–244 (2003).

Morf, H., “The stochastic two-state solar irradiance model (STSIM),” Solar Energy **62**, 101–112 (1998).

Morf, H., “A stochastic solar irradiance model adjusted on the Ångström–Prescott regression,” Solar Energy **87**, 1–21 (2013).

Munkhammar, J. and Widén, J., "Copula correlation modeling of aggregate solar irradiance in spatial networks," *Proceedings of the 6th International Workshop on Integration of Solar Power into Power Systems*, Vienna, Austria, 14–16 November 2016.

Munkhammar, J. and Widén, J., "An autocorrelation-based copula model for generating realistic clear-sky index time-series," Solar Energy **158**, 9–19 (2017a).

Munkhammar, J. and Widén, J., "An autocorrelation-based copula model for producing realistic clear-sky index and photovoltaic power generation time-series," *Proceedings of the IEEE PVSC-44*, Washington DC, USA, 25–30 June 2017b.

Munkhammar, J. and Widén, J., "A Markov-chain probability distribution mixture approach to the clear-sky index," Solar Energy **170**, 174–183 (2018a).

Munkhammar, J. and Widén, J., "An N-state Markov-chain mixture distribution model of the clear-sky index," Solar Energy **173**, 487–495 (2018b).

Munkhammar, J. and Widén, J., "A spatiotemporal Markov-chain mixture distribution model of the clear-sky index," Solar Energy **179**, 398–409 (2019).

Murphy, K. P., *Machine Learning: A Probabilistic Perspective* (MIT Press, Cambridge, MA, 2012).

Mustacchi, C., Cena, V., and Rocchi, M., "Stochastic simulation of hourly global radiation sequences," Solar Energy **23**, 47–51 (1979).

Nelsen, R. B., *An Introduction to Copulas*, 2nd ed., Springer Series in Statistics (Springer, New York, 2006).

Ngoko, B. O., Sugihara, H., and Funaki, T., "Synthetic generation of high temporal resolution solar radiation data using Markov models," Solar Energy **103**, 160–170 (2014).

Peruchena, C. F., Larrañeta, M., Blanco, M., and Bernardos, A., "High frequency generation of coupled GHI and DNI based on clustered dynamic paths," Solar Energy **159**, 453–457 (2018).

Poggi, P., Notton, G., Muselli, M., and Louche, A., "Stochastic study of hourly total solar radiation in Corsica using a Markov model," Int. J. Climatol. **20**, 1843–1860 (2000).

Sengupta, M. and Andreas, A. "Oahu solar measurement grid (1-year archive): 1-second solar irradiance," National Renewable Energy Laboratory, Oahu, Hawaii (Data), https://data.nrel.gov/submissions/11 (accessed 29 August 2020).

Shepero, M., Widén, J., and Munkhammar, J., "A generative hidden Markov model of the clear-sky index," J. Renew. Sustain. Energy **11**, 043703 (2019).

Shumway, R. H. and Stoffer, D. S., *Time Series Analysis and Its Applications: With R Examples*, 3rd ed. (Springer, New York, 2011).

Smith, C. J., Bright, J. M., and Crook, R., "Cloud cover effect of clear-sky index distributions and differences between human and automatic cloud observations," Solar Energy **144**, 10–21 (2017).

Suehrcke, H. and McCormick, P. G., "The distribution of average instantaneous terrestrial solar radiation over the day," Solar Energy **42**, 303–309 (1989).

Sun, X., Bright, J. M., Gueymard, C. A., Acord, B., Wang, P., and Engerer, N. A., "Worldwide performance assessment of 75 global clear-sky irradiance models using principal component analysis," Renew. Sustain. Energy Rev. **111**, 550–570 (2019).

Sun, X., Bright, J. M., Gueymard, C. A., Bai, X., Acord, B., and Wang, P., "Worldwide performance assessment of 95 direct and diffuse clear-sky irradiance models using principal component analysis," Renew. Sustain. Energy Rev. **135**, 110087 (2021).

Vergara-Dominguez, L., Garcia-Gomez, R., Figueiras-Vidal, A. R., Casar-Corredera, J. R., and Casajus-Quiros, F. J., "Automatic modelling and simulation of daily global solar radiation series," Solar Energy **35**, 483–489 (1985).

Vretblad, A., *Fourier Analysis and Its Applications*, Graduate Texts in Mathematics (Springer-Verlag, New York, 2003).

Wegener, J., Lave, M., Luoma, J., and Kleissl, J., "Temporal downscaling of irradiance data via hidden Markov models on wavelet coefficients: Application to California solar initiative data," University of California San Diego, 2012.

Widén, J. and Munkhammar, J., *Solar Radiation Theory* (Uppsala University, Uppsala, 2019a).

Widén, J. and Munkhammar, J., "Spatio-temporal downscaling of hourly solar irradiance data using Gaussian copulas," *Proceedings of the 46th IEEE Photovoltaic Specialists Conference (PVSC)*, Chicago, USA, 16–21 June 2019b.

Widén, J., Shepero, M., and Munkhammar, J., "On the properties of aggregate clear-sky index distributions and an improved model for spatially correlated instantaneous solar irradiance," Solar Energy **157**, 566–580 (2017).

Zhang, W., Kleiber, W., Florita, A. R., Hodge, B. M., and Mather, B., "A stochastic downscaling approach for generating high-frequency solar irradiance scenarios," Solar Energy **176**, 370–379 (2018a).

Zhang, W., Kleiber, W., Florita, A. R., Hodge, B. M., and Mather, B., "Modeling and simulation of high-frequency solar irradiance," IEEE J. Photovoltaics **9**, 124–131 (2018b).