Solar Energy 182 (2019) 462–479

Contents lists available at ScienceDirect

Solar Energy

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

Nonparametric Bayesian-based recognition of solar irradiance conditions: Application to the generation of high temporal resolution synthetic solar irradiance data

Âzeddine Frimane[a,*], Ted Soubdhan[b], Jamie M. Bright[c], Mohammed Aggour[a]

[a] Laboratory of Renewable Energies and Environment (LR2E), Faculty of Science, Ibn Tofail University, B.P 133-14000 Kenitra, Morocco
[b] Laboratoire LARGE, Université des Antilles, 97157 Pointe à Pitre cedex, France
[c] Fenner School of Environment and Society, The Australian National University, 2601 Canberra, Australia

ARTICLE INFO

Keywords:
Solar irradiance
Clustering
Clearness index
Bayesian nonparametric
Synthetic irradiance

ABSTRACT

High resolution synthetic irradiance is of interest for theoretical studies such as grid integration of solar PV and battery storage analysis. Access to site-specific data is often limited to inadequate temporal resolutions for such application. A new model for producing synthetic solar global horizontal irradiance (GHI) time-series at up to 1-min resolution is presented as derived from >10-min input data. Briefly, it is a clustered-based method for daily clearness index distributions using Dirichlet process Gaussian mixture model (DPGMM). DPGMM is a non-parametric Bayesian (NPB) model indexed with an infinite-dimensional space of parameters. The key benefit of the NPB paradigm is the automatic adaptation to the correct complexity level and model size, suggesting a local adaptation of the model to all climatic conditions. A posterior inference using Markov chain Monte Carlo algorithm (namely Gibbs sampling) is applied. The model only requires a valid number of intraday data to construct daily distributions, then it can be applied worldwide. The synthetic GHI time series are validated against observed 1-min GHI data for four locations distributed throughout the world with different climatic conditions and significant geographic separation. Moreover, the presented method can generate data based on similar climatic conditions. A good fit between real and generated data is observed. We present an nRMSE $\leq 4\%$ and nMBE $< \pm 4\%$ between generated and measured means at both daily and monthly scales for all sites. The agreement between the real and generated cumulative density distributions of six comparative variability metrics (defined in text) at four different sites is measured using the overlapping and the Kullback–Leibler coefficients, which are $\geq 75\%$ and $\leq 10\%$ respectively, in all cases. To ensure the reproducibility of the research presented in this paper, the methodology is freely available as an R-package downloadable from SolarClusGnr.

1. Introduction

Solar irradiance at the top of the atmosphere is highly predictable and can be estimated using the solar constant (1361.1 Wm^{-2}, (Gueymard, 2018)) and Sun-Earth distance. On the ground, however, it can be modelled as a stochastic process due to multiple factors including solar geometry, cloud distribution, and various other extinction processes (Lave et al., 2015). As a result, the input power of solar energy systems (SES) is intermittent in nature and does not ensure optimal network operation. This poses a number of problems for grid operators, such as deficiency compensation, power quality and stability issues, mainly in small or not interconnected electrical networks as found in islands (Yang et al., 2012; Schallenberg-Rodríguez and Montesdeoca, 2018; Bright et al., 2017). For this reason, a robust identification of the solar irradiance variability is a crucial step to guide a successful integration of SES, including thermal and photovoltaic (PV) processes.

1.1. An overview of literature

Classification strategies for solar irradiance are thought to be an effective solution for investigating the behavior of such stochastic processes. It allows the inferring of important information from data that can represent a preliminary facilitating level for further processing, such as solar irradiance forecasting (Ghayekhloo et al., 2015) or synthetic irradiance time series generation as suggested in this paper. Classification is a maturing field of machine learning, which is now being spun out into renewable energy applications in general and solar energy in particular. A comprehensive review of the common

* Corresponding author.
E-mail address: Azeddine.frimane@uit.ac.ma (Â. Frimane).

https://doi.org/10.1016/j.solener.2019.02.052
Received 16 December 2018; Received in revised form 18 February 2019; Accepted 20 February 2019
Available online 07 March 2019
0038-092X/

classification techniques used for solar irradiation and other alternative renewable energy sources can be found in Pérez-Ortiz et al. (2016). For instance, Moreno-Tejera et al. (2017) used the k-medoids algorithm to cluster days as a function of the sky state for concentrated solar power operation. Three indexes are used in this analysis, the transmittance index, persistence index of the instantaneous transmittance index values and variability index. Munshi and Mohamed (2016) apply a set of clustering methods from different clustering categories to determine the optimum number of clusters for photovoltaic power patterns data. Zagouras et al. (2014) investigated the development of maps created by the combination of two well-known clustering techniques; i.e, the affinity propagation and the k-means. This methodology makes it possible to select candidate locations for solar power plants, to determine regions of coherent solar quality attributes and to improve solar forecasting for PV plants. Kang and Tam (2013) deal with classification of daily sky conditions by using the daily clearness index and the daily probability of persistence. Support vector machine (SVM) has also often been used for solar irradiance classification. For example, it is applied by Lee et al. (2004) to the detection and classification of clouds for use in earth observing system models. Three classes were recognized in this work: clear sky, water cloud and ice cloud. In Wang et al. (2015) a SVM based weather status pattern recognition is proposed for a short-term PV power prediction.

Classification models can be organized according to several aspects, mainly depending on the learning approach (supervised or unsupervised) and the nature of the model assumptions (parametric or nonparametric). The correct classification is always unknown and its interpretation varies from model to model. This raised a number of questions, notably about how to select the optimal number of classes and the correct model complexity. Previous studies related to solar irradiance time-series classification have mostly focused on parametric models which manipulate data with a priori fixed model complexity and size. However, in most cases, the prior information will be insufficient to justify these parametric assumptions such as distance measure, density threshold, ···, which do not allow their application without major alterations (Rasmussen, 1999). Moreover, the unknown number of solar irradiance classes must be specified in advance according to a chosen indicator which can result in the over-fitting effect if an inappropriate index is selected (to our knowledge, about 30 have been published in the literature). Another important issue is that parametric methods will return a partition even if data do not contain any structure. Table 1 shows the resulting clustering using three well-known parametric methods in the solar energy literature, namely the K-means, the partition around medoids (PAM) and hierarchical clustering (HC). In this example, we use 10-min averaged set of Sioux Falls, USA, as a training set (See Section 2 for more details about data). To choose the optimal number of classes, we maximize the silhouette score and Dunn score (Brock et al., 2008) under different distance measures for the same algorithm. Table 1 clearly shows that the number of classes varies from model to model, as such, the correct number of classes is highly subjective. It depends on the algorithm used and the measuring parameter of similarity, which makes the appropriate choice ambiguous.

Table 1
Variation of the number of classes using the K-means, the partition around medoids (PAM) and hierarchical clustering (HC) under different distance measures, maximizing the silhouette score and Dunn score.

	Euclidean distance	Mahalanobis distance	Maximum distance	Manhattan distance
	Dunn score			
K-means	4	4	5	2
PAM	2	3	2	2
HC	3	5	3	3
	Silhouette score			
K-means	2	2	5	2
PAM	2	2	4	2
HC	2	2	2	2

One of the optimal approaches to solve these problems involves the use of the nonparametric paradigm. It is able to establish automatically the correct number of classes and even able to detect the absence of partition in the data structure, without having to perform any model comparisons (Gershman and Blei, 2012). The nonparametric method prescribes the use of an infinite-dimensional parameter space rather than a fixed dimension. Furthermore, a common practice is to combine the Bayesian approach with the nonparametric framework to provide more robust models based on the prior knowledge (Frimane et al., 2018), which can be of interest for different climates and geographical differences at various sites around the world.

Traditionally, SES studies and simulations are primarily conducted on the basis of historical solar irradiance measurements or through public databases such as typical meteorological years. It is also possible to use measured data on sites close to the place of use. The key problem with such data is that it is often available with an unsatisfactory time resolution and not representative of the actual solar irradiance variability for relevant time scales (1-min resolution or better) (Lave et al., 2015). Therefore, these data cannot be used to determine the impacts of PV, such as how distributed generation of PV impacts the low voltage electricity grid (Mateo et al., 2018). Using data with an inappropriate characteristics would result in an inappropriate SES integration. From another side, satellite imagery can offer near-global coverage of solar irradiance data, the Americas and the Pacific regions recently have temporal resolutions down to 10-min (Engerer et al., 2017). Whilst the spatial coverage is excellent, it is still not high-enough resolution without sufficient downscaling (Bright et al., 2018).

An alternative approach to obtaining high-resolution irradiance data is through synthetic irradiance generation. A large body of models have been proposed in literature, they are predominantly temporal only methodologies with some notable spatio-temporal models (Bright et al., 2017; Bright et al., 2015; Bright, 2019; Grantham et al., 2017; Ngoko et al., 2014; Munkhammar et al., 2017; Munkhammar and Widén, 2016, 2018, 2019; Peruchena et al., 2018, 2015; Zhang et al., 2018; Shi et al., 2018; Polo et al., 2011; Larrañeta et al., 2018). For instance, Peruchena et al. (2018) presented a clustered-based methodology for the generation of 1-min coupled global horizontal irradiance (GHI) and direct normal irradiance (DNI) temporal time series on the basis of the envelope clear sky and Dynamic Paths concepts. Larrañeta et al. (2018) proposed a geographically flexible methodology to synthetically downscale DNI time-series from 1-h to 1-min temporal resolution. Grantham et al. (2017) developed a method for generating 5-min resolution temporal-only synthetic time series of GHI and DNI from stored sequences guided by hourly average values. This approach was advanced further in a temporal only 1-min resolution, downscaling, synthetic time series generation of GHI through a Markov chain approach by Bright (2019). Observed sequences of GHI were binned according to the mean hourly zenith and clear-sky index. By testing on different pairs of training and testing, they concluded that variability clustering would be the most likely cause of similarity between two sites. Bright et al. (2015) present a method for synthetically generate 1-min temporal GHI from hourly weather observations including sea level pressure, wind speed, cloud base height and cloud cover. The method employs Markov transition matrix to determine the cloud cover index. Later, the method was improved by Bright et al. (2017) by adding spatial correlation in the synthetic generation procedure resulting in a spatially decorrelating solar irradiance generator. Ngoko et al. (2014) present a model for 1-min global temporal solar irradiance generation using Markov chain model and starting by grouping daily clearness index based on predefined thresholds. In Zhang et al. (2018), prior to the data generation process, the method involves an initial decision tree to determine the

state of the day: for a given day, if the maximum of the first order differences is less than a predefined threshold (set to 0.1) the day is considered clear otherwise it is non-clear. In our case, this is guaranteed by the NPB clustering paradigm. It avoids the predefinition of such thresholds and any other parametric assumptions such as the number of day states. All parameters within the NPB are directly inferred from data and so cannot be biased by predefinition—it allows the data speak for itself. A spatial only model for synthetic irradiance is presented by Munkhammar and Widén (2016). They achieve modelling the spatial dimension by calculating the cross correlation of the clear-sky index between different stations within a certain spatial domain are using the copula to guarantee the correlation is maintained. A temporal only model (Munkhammar and Widén, 2018) used Markov chains to model the state of clear-sky index and how it would probabilistically change over time. When the aforementioned spatial-only and temporal-only methodologies were combined, a spatio-temporal Markov-chain mixture distribution model was presented by Munkhammar and Widén (2019). They used the clear-sky index at an arbitrary number of locations whilst maintaining the cross-correlation to generate spatio-temporal synthetic irradiance profiles. A rich literature review of the topic is presented by Munkhammar and Widén (2019). We do not present a spatial methodology in this work at this time, however, it is expected that a combination of the NPB temporal approach of the presented work with the spatial methodology by Munkhammar and Widén (2016), a spatial element could be introduced at a later date.

A significant observation from these studies in our overview of literature is that many of them ignore the difference between day types. Different types of day have different statistical properties, treating them in the same manner can bias the results. A good example of this is in Bright et al. (2017) whereby the daily variability index is not satisfactorily captured as days that are very clear are synthetically reproduced with significant ramp events due to the nature of the methodology. While some of them start by grouping data, they utilize traditional parametric methods that suffer from what has been discussed previously. Inconsistent classification produces different results for the same data at different time scales which is not suitable for discovering the real inherent data structure. Furthermore, some of them require a very repetitive and inefficient operation of sequence storage (e.g. Grantham et al. (2017)) and others require the use of a long period of observational data for model training, as well as meteorological data not easily obtained (e.g. Bright et al. (2017)). Furthermore and from the methodological point of view, they are often not accompanied by a well-documented and easy-to-adopt code base, which renders them highly-unlikely to be reused.

1.2. Main ideas and contribution

This paper presents a new clustering-based methodology for generation of temporal 1-min synthetic irradiance data as downscaled from 10-min to 20-min real data such is the standard availability from next generation satellite databases. We state that the method proposed within is not limited to GHI, though this is the only irradiance component that it is applied to, but could also be applied to direct and diffuse irradiance also. The new method aims to satisfy all the key opportunities identified from literature. Firstly, a Dirichlet process Gaussian mixture model (DPGMM) is employed that aims to provide a new consistent nonparametric Bayesian (NPB) framework for automatic classification, commonly referred to as clustering of daily clearness index distributions. The DPGMM investigates the use of an overarching probability function; i.e, the multivariate Gaussian distribution (MGD), combined with the NPB paradigm. This allows inferring robust conclusions only from data (free from parametric assumptions—bias) and avoids all previously cited issues with the parametric models. The main advantage of using the MGD come from the fact that it is a more flexible in terms of co-variance, and thus it can be rotated, scaled and adapted easily. Moreover, it does not depend on any intrinsic properties of the measurements, then, it does not suffer from the samples size alignment or from the sampling rate compared to the multinomial distribution for example in Frimane et al. (2018). Mathematically, the key benefit of the MGD choice over other distributions is that it can represent a valuable estimate of the data distribution because of the central limit theorem. Also, it presents an attractive pattern due to its computational straightforwardness and its ease of interpretation.

The main purpose of the DPGMM is to automatically divide the days into similar groups that share common characteristics. Each group corresponds to a specific meteorological regime, described statistically by its corresponding component in the mixture distribution. The NPB nature of the model is aimed at minimizing errors and ensuring the closeness of the number and sequence of the resulting classes at different sampling rates of the solar irradiation data. Therefore, the resulting mixture distribution will be able to generate synthetic irradiance data regardless of the original temporal resolution of the solar irradiance signal. To ensure that the simulated occurrence order of intraday data has the same order distribution as the actual data, we use the Markov chain model. For each cluster, the clearness index time-series is concatenated and discretized into a sufficient number of states to infer the underlying transition matrix and thus offers the possibility to predict of states dynamic in this cluster.

The proposed method is validated in a number of sites with different climatic conditions and latitudes around the world, suggesting their application worldwide. The method is also able to generate synthetic data based on available similar climatic clustering. Moreover, to compare real and generated data, we use a recently published set of quantifiers (Blaga and Paulescu, 2018) that better highlight the different facets of solar irradiance variability. The overlap coefficient (OVC) (Inman and Jr, 1989) also called Szymkiewicz-Simpson coefficient, and Kullback–Leibler divergence (KLD) (Kullback and Leibler, 1951) called also the relative entropy measure in literature are used to quantify the good-fit between measured and simulated data. From the computational viewpoint, our algorithm requires less input data (just a time-series of solar irradiance) and is more computationally straightforward since it is Bayesian. The computation task is based on the generation of drawings from the posterior distributions using Markov chain Monte Carlo procedure, specifically Gibbs sampling (Neal, 2000).

Most importantly, reproducibility of this work can be achieved by downloading the R-package methodology from https://github.com/frimane/SolarClusGnr.

The structure of this paper is as follows. Section 2 provides the description of the solar irradiation databases used in this study. The methodology is fully detailed in Section 3. Sections 4.1 and 4.2 discuss the experimental results of classification and presents a detailed analysis of the resulting classes and class sequences respectively. The results of the solar irradiance time-series synthesis and the model weaknesses are presented in Section 5. Closing remarks are presented in Section 6. As a complement to this article, we provide the explicit expression of the used probability functions in Appendix A and some examples of simulated days in Appendix B.

2. Solar global horizontal irradiance data

Four data-sets of GHI time-series distributed throughout to the world were used to validate the method described in this paper. All irradiance data is gathered from the Baseline Surface Radiation Network (BSRN) an affiliate of the World Meteorological Organization (WMO). All the irradiance data collected underwent rigorous quality control following the well established convention described by Long and Shi (2006). These checks test the measured values against their physical limits for acceptability. Measured GHI data from the BSRN is not taken directly from the measured instrument directly, but instead calculated as the sum of the direct horizontal irradiance and diffuse horizontal irradiance as is the convention in solar engineering Gueymard and Myers (2009). This is due to the increased accuracy of

Table 2
Characteristics of the solar irradiation data-sets and corresponding Köppen climate classification.

Location	Abbreviation	Latitude	Longitude	Time span	Köppen climate classification
Alice Springs,Australia	ASP	23.80°S	133.89°W	2014–2015	Arid Desert Hot (BWh)
Tamanrasset, Algeria	TAM	22.79°N	5.53°E	2016–2017	Arid Desert Hot (BWh)
Toravere, Estonia	TOR	58.25°N	26.46°E	2015–2016	Humid Continental (Dfb)
Sioux Falls, USA	SXF	43.73°N	96.62°W	2015–2016	Humid Continental (Dfb)

Fig. 1. The geographical locations of the four selected BSRN stations built using R-worldmap©(Kahle and Wickham, 2013).

the derived product, particularly at low sun angles due to reduced determination uncertainty and cosine responses Michalsky et al. (1999). The BSRN use Kipp & Zonen CM11 pyranometers for diffuse horizontal irradiance and CH1 pyrheliometers for direct normal irradiance. The geographical coordinates, abbreviations of location names used in this paper, time span of the measurements and the Köppen climate classification are summarized in Table 2. Fig. 1 clearly shows the spread out of the meteorological stations throughout the world.

The Arid Desert Hot climate (BWh) is the most common climate type by land area (14.2%). The cloud cover in this climate is uncommon, rendering solar irradiance under these conditions more stable. It is also characterized by high temperature throughout the year and can be subject to rapid changes in aerosol which can dramatically influence the direct and diffuse irradiance. Humid Continental climate (Dfb) is characterized by large differences in seasonal temperature and rarely have extremely high irradiance. Readers can find more interesting information on Köppen climate classification in (Peel et al., 2007).

3. Methods

Importantly, the methodology presented is not limited to solar irradiance, as such, we have presented the mathematical formulation in a more generic way such that it can be reproduced to any other application. That said, we first describe the methodology directly in terms of GHI for conceptual understanding before disclosing the more complex nature of the composition itself. We have tried to draw GHI specific examples and analogies throughout to assist with understanding.

In overview, the DPGMM can be separated into four key sections. Fig. 2 summarizes the methodology showing how it is segmented into rectangles each representing a separate procedure discussed in an independent subsection. Section 3.1 discusses the time series style data that is required in our application of the DPGMM and how the data is condensed into distributions so that it can be used in data synthesis. In our application, our input data is GHI that is converted to clearness-index and represented as distributions from a full day's data. The next stage of the DPGMM is the prior distributions setting discussed in Section 3.2. 'Prior distributions' or simply 'priors' represent the base information about the structure of the classes before taken into account the data. They help define how different each daily distribution is so that an appropriate mixture of significantly unique classes can be detected. We talk about a 'class' regularly in this methodology. Explicitly, we are talking about the mathematical classification of each daily distribution in terms of the probability likelihood, e.g. how likely is it that this day's distribution belongs to a class. We might label these classes with descriptive terms that would change depending on the application so that they are easier to conceptually understand. In the GHI application, a class is a essentially a description of the type of day i.e. clear, cloudy, intermittent-cloud, overcast etc.

Next, the Posterior inferences must be established (Section 3.3). 'Posterior probabilities' or just 'posteriors' represent the updated priors after the data has been considered—much like a collection of daily distributions that are all separated into their appropriate classes which are ready to be used for synthetic generation. The DPGMM can be generically applied to any number of classifications that are inherently contained within the data. In summary, the posterior considers each daily distribution in turn. Each daily distribution is compared to all other previously analyzed distributions in order to define the class of new daily distribution; the result is classification into a pre-existing class, or to form a new class entirely. This approach means that we can separate the data into *all* possible classes that exist in the data. In the GHI application, this means that every single site is analyzed for all possible classes. Consider a Saharan climate, we might expect to see the majority of days being clear, with some intermittent days and very few overcast days— the DPGMM may only result in three classes. Then consider a highly variable climate such as high latitudes or tropical, we expect far more varied weather conditions that will result in different types of day (clear, intermittently-clear, broken clouds, slightly overcast, heavily overcast, partially cloudy etc.). Hence, this ability to adapt to the data without pre-definition is a significant strength of the DPGMM.

Lastly, we bring together the methodology above to reproduce synthetic time series in Section 3.4. This section demonstrates how time series can be synthetically generated whilst maintaining the class distribution and probabilistic nature of the time series. In our application, we are synthesizing GHI data in a stochastic manner. The data, R-package and code-example of all methods used in this paper—clustering and data generation, are made available in the supplementary materials at Appendix C.

3.1. Data representation

The DPGMM is a multivariate model that is highly versatile and can be applied to many types of time series data. Any data variable that can be summarized by a regular distribution over a fixed temporal duration (e.g. every day/month/year) can be synthetically generated using the DPGMM. Whilst our chosen application is to use GHI data, we could also have applied this to any synthetic time series production e.g. temperature, wind speed, currency value, house prices within a city, petroleum prices. This is not to say synthetic time series of all these applications would be useful, it is merely to reiterate that the DPGMM methodology can operate on time series that can be classified over time. Furthermore, the methodology is not strictly limited to single vector time series data, it could also be applied to three-dimensional variables such as the 3D-wind speed vector, or could consider multiple correlated (or uncorrelated) variables so long as the distributions and correlation believes are set in the prior distribution settings to facilitate representation.

In the literature, several dimensionless indices have been used to characterize the transparency of the atmosphere. The most common one is the clearness index k_t (Perez et al., 1990). It is used because it is simple to calculate and avoid the uncertainty introduced by atmospheric variables

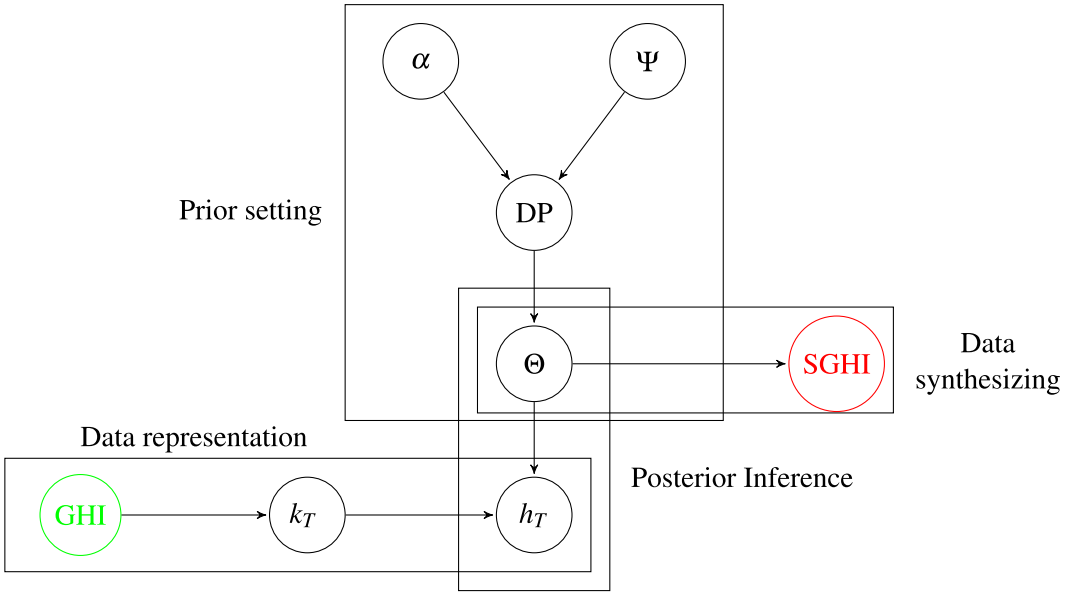

Fig. 2. Graphical representation of the methodology: Each rectangle represents a procedure. Firstly, data is pre-processed in the rectangle 'Data representation' where GHI is the global horizontal irradiance, k_t is the daily clearness index and h_t is the daily clearness index distribution. After priors setting, the process of clustering is initialized to infer cluster characteristics Θ. Dirichlet process (DP) is the prior on Θ where α is the concentration parameter of DP, and Ψ is the base distribution. Based on Θ, high resolution global horizontal irradiance data (SGHI) are synthesized.

and complex clear-sky irradiance modeling Zhong and Kleissl (2015). The clearness index is defined as $k_t = \frac{E_g}{E_0}$, where E_g is the estimated GHI at the ground and E_0 is the horizontal extraterrestrial irradiance. In this work, we prefer to use k_t instead of the normalized clearness index k_t'. The latter is defined as a zenith angle independent index, normalizing k_t with respect to a clear sky global irradiance profile (Perez et al., 1990). k_t' has the tendency to gather data around the mean, which is not desireable for our classification method, in particular, that we rely on the mode(s) of k_t distribution(s) to characterize the classes.

For the data preprocessing stage, k_t is represented as daily equal width histograms $(h_t)_{t=1,2,\ldots,N}$ (N is the total number of days) based on a fixed partition of the range [0–1], which are considered as nonparametric density estimators. In the absence of a complete characterization of how many bins should be used in a unified way for all days (García et al., 2013; Sun et al., 2009), we must find an appropriate method that satisfies two criteria: the distributions h_t must (i) all have same bin widths, (ii) not contain empty bins in the distribution. Binning data is difficult. We reviewed the many options available in Section 1 and opted to follow the method used by Soubdhan et al. (2016) where the number of bins selected provides a valid shape of h_t without leading to empty bins. With sample rates longer than 10–20 min, h_t will not provide useful information because there will be too few data per day to build a distribution for a day (Smith et al., 2017). This is particularly true at higher latitudes where days can be very short. Hourly sampling rates would regularly result in fewer than 8 samples, which is insufficient to build a meaningful distribution.

3.2. Prior setting

This section describes the prior settings around which the DPGMM is based. All quantities within the DPGMM (e.g. k_t, parameters, hyperparameters, etc.) are random variables, and so they are represented by distributions. We define the distribution of k_t by allowing an infinite number of Gaussian distributions that, when combined, form the overall distribution. We adopt a Bayesian approach and so we must establish a prior set of parameters that define these infinite number of Gaussian distributions before they are later updated in the posterior. Prior settings provide the first description of the classes that exist within the data before it takes into account any input data or other evidence.

Because the parameters that define the overall k_t distribution are also random variables, they in turn are represented by their own set of parameters— we call these additional sets of parameters 'hyperparameters'. Hyperparameters are again random variables, however, we do not extend the hierarchy beyond this level of complexity; the hyperparameters are fixed by the empirical mean and the covariance of the input data. The explicit expression and parameterization of each probability function used in this section is listed in Appendix A.

For the DPGMM, the prior setting begins by having an infinite number of classes. What this means is that every dataset analysed by this methodology does not have a limit on the number of classes identified as statistically. Therefore, each daily distribution of k_t (h_t) is generated by an infinite mixture of multivariate Gaussian distributions (MGDs). These concepts are denoted in what follows:

$$h_t|\{\pi_j, \mu_j, \Lambda_j\}_{j=1}^{\infty} \sim \sum_{j=1}^{\infty} \pi_j \mathcal{N}_j(h_t|\mu_j, \Lambda_j), \tag{1}$$

where $\pi = \{\pi_j\}_{j=1}^{\infty}$ is the mixing proportions vector, $\Theta = \{\mu_j, \Lambda_j\}_{j=1}^{\infty}$ are the parameters of the MGDs components; μ_j is the mean vector for the MGD component $\mathcal{N}_j$ and Λ_j its precision matrix. The use of the precision matrix is the author's preference and not an obligation of the model. An alternative approach would be to use a covariance matrix, however, the result is the same. The symbol ~ is read proportional to.

Dirichlet process (DP) (Ferguson, 1973) is one of the most used nonparametric priors. It is a random generator of probability distributions, defined by a base distribution and a concentration parameter. If the prior and posterior distributions belong to the same family, they are called conjugate distributions. The prior in this case is called conjugate prior. When the conjugate prior is multiplied by the proper likelihood function, the result is a closed-form expression to the posterior distribution. To achieve the conjugation in our modelling, we follow the choice of the base distribution Ψ introduced by Görür and Edward Rasmussen (2010). The prior of μ_j is also an MGD denoted as $\mathcal{N}$ conditioned on Λ_j and the prior of Λ_j is a Wishart distribution denoted as $\mathcal{W}$:

$$\mu_j|\Lambda_j, \xi, \rho \sim \mathcal{N}(\mu_j|\xi, (\rho\Lambda_j)^{-1}) \tag{2a}$$

$$\Lambda_j|\nu, W \sim \mathcal{W}(\Lambda_j|\nu, (\nu W)^{-1}), \tag{2b}$$

then, the base distribution for the DP is Normal-Wishart distribution denoted as follows:

$$\Psi \sim \mathcal{NW}(\xi, \rho, \nu, W), \tag{3}$$

where the hyperparameters variables ξ, ρ, ν and W are distributed as:

$$\xi \sim \mathcal{N}(\xi|\mu_h, \Lambda_h^{-1}) \tag{4a}$$

$$\rho \sim Gamma(\rho|1, 1) \tag{4b}$$

$$(\nu - D - 1)^{-1} \sim Gamma((\nu - D - 1)^{-1}|1, D^{-1}) \tag{4c}$$

$$W \sim \mathcal{W}(W|D, (D\Lambda_h)^{-1}), \tag{4d}$$

where μ_h and Λ_h are respectively: the empirical mean vector and the precision matrix of our h_t data-set. *Gamma* is the Gamma distribution and D is the number of bins of h_t. Note that μ_h and Λ_h subscripted by h are not functions of h_t and this is to indicate that they are different from μ and Λ defined before as the mean and precision of the Gaussian components of the mixture model. The hyperparameters are the parameters of the base distribution of DP. They are common to all components and represent our beliefs where the class parameters are similar and represent the same class. This physically translates our expectation of the differences in deviations and means that define distinct GHI classes. The base distribution Ψ encapsulates all our prior knowledge of the chosen likelihood functions for data.

Görür and Edward Rasmussen (2010) proposed a Gamma prior for the inverse of the concentration parameter α^{-1}:

$$\alpha^{-1} \sim Gamma(\alpha^{-1}|1, 1). \tag{5}$$

The concentration parameter controls the relative contribution of the prior and data to the posterior distribution.

From the nonparametric viewpoint, the number of mixture components is infinite. As a result, an infinite number of variables must be drawn from Ψ and, therefore, all normalization processes must fail. In practice, the stick-breaking representation introduced by Sethuraman (1994) is used to avoid such problems and thus sample from the DP. The complete model can be written now as follow:

$$\begin{aligned} \pi|\alpha &\sim GEM(\pi|\alpha) \\ (\mu_c, \Lambda_c) &\sim \Psi \\ c_t|\pi &\sim Categorical(c_t|\pi) \\ h_t|c_t, \mu_{c_t}, \Lambda_{c_t} &\sim \mathcal{N}(h_t|\mu_{c_t}, \Lambda_{c_t}), \end{aligned} \tag{6}$$

where $GEM(\alpha)$ is the distribution of weights from the stick-breaking process and $c = \{c_t\}_{t=1}^N$ are the indicator variables which indicate classes associated with observations. $t = 1, 2, ...,N$ and N is the total number of days.

3.3. Posterior inference

In the posterior inference, we decide whether the observation under analysis belongs to an existing class or whether a new class must be created. In our application, we are concerned with the distribution of k_t from an entire day. Our prior settings are therefore updated with the data to determine which class the k_t distribution belongs. We can arbitrarily name these classes depending on their distribution characteristics, e.g. a distribution with a very high peak at a large k_t would be indicative of a clear day.

The prior distribution of c_t—the prior distribution of the class label that the day t takes— can be obtained by integrating out the mixing proportions π conditional on all other random variables. It has the following limits at infinity. For more details, see Neal (2000):

$$\mathcal{P}(c_t = j; j \text{ already seen}|\ c_{-t}, \alpha) \longrightarrow \frac{n_{-t,j}}{N - 1 + \alpha}, \tag{7a}$$

$$\mathcal{P}(c_t = j; j \text{ new}|\ c_{-t}, \alpha) \longrightarrow \frac{\alpha}{N - 1 + \alpha}, \tag{7b}$$

where c_{-t} is the set of all observed indicator variables except c_t, $n_{-t,j}$ is the number of days associated with the class label j excluding the day t, and N is the total number of days.

By multiplying the Gaussian likelihood MGD of the GHI classes by the prior distributions of Eq. (7a) and integrating it over Eq. (7b), we obtain the posterior probability that the day t belongs to an existing class and the probability of creating a new class for the day t, respectively. Consider the scenario where the DPGMM is analysing the day t and has so far already identified two classes from the data: clear and intermittent-clear. Should t have significantly distinct properties that do not satisfy clear and intermittent-clear, a third class is instead produced—this new class might represent cloudy/overcast. Thus, Gibbs sampling for the indicator variables is based on the following conditional posterior probabilities:

$$\mathcal{P}(c_t = j; j \text{ already seen}|c_{-t}, h_t, \mu_j, \Lambda_j, \alpha) \propto \frac{n_{-t,j}}{N - 1 + \alpha}\mathcal{N}(h_t|\mu_j, \Lambda_j). \tag{8a}$$

$$\mathcal{P}(c_t = j; j \text{ new empty class}|\ c_{-t}, h_t, \alpha, \xi, \rho, \nu, W) \propto \frac{\alpha}{N - 1 + \alpha} \times \iint \mathcal{N}(h_t|\mu, \Lambda)... \mathcal{NW}(\mu, \Lambda|\xi, \rho, \nu, W)d(\mu)d(\Lambda), \tag{8b}$$

Note that, the symbol $\propto$ is read 'proportional to'.

Analytic computation of posterior distributions is intractable (Neal, 2000), meaning that it cannot be solved analytically and so computer computation is required, namely Gibbs sampler. Algorithm 1 summarize Gibss sampling of the DPGMM.

Algorithm 1.

Data: h_t
Result: clustered h_t
Summarize k_t as daily equal width histogram h_t;
Randomly initialize c_t for each h_t;
while $i <$ *number of iteration* **do**
  Update ξ, ρ, ν and W;
  Update $\{\mu_j, \Lambda_j\}_{j=1}^k$;
  for j *in* $1 : N$ **do**
    Remove the old assignment c_t for h_t;
    if *Its class being empty* **then**
      Remove it and decrease the number of classes;
    end
    for *each component* **do**
      Update c_t, conditional on c_{-t}, $\{\mu_j, \Lambda_j\}_{j=1}^k$, ξ, ρ, ν and W;
    end
    Update α;
  end
end

3.4. Generation of synthetic time series

Now that the data has been analyzed, the prior and posterior distributions are realized and the Markov chain Monte Carlo technique facilitates the transitioning of classes, it is fully possible to generate a synthetic time series.The method is fully described in the Algorithm 2 below. The time series produced considers only the temporal dimension and cannot consider appropriately correlated spatial dimensions.

In our application, the main point considered in the construction of the synthetic solar irradiance time-series is the consistency of the DPGMM at different temporal resolutions of data. DPGMM being consistent means that the number and sequence of the resulting classes of two time-series are similar if and only if the mixture of distributions that generated them are similar. Accordingly, DPGMM is able to explain the data with similar mixtures of distributions by revealing similar

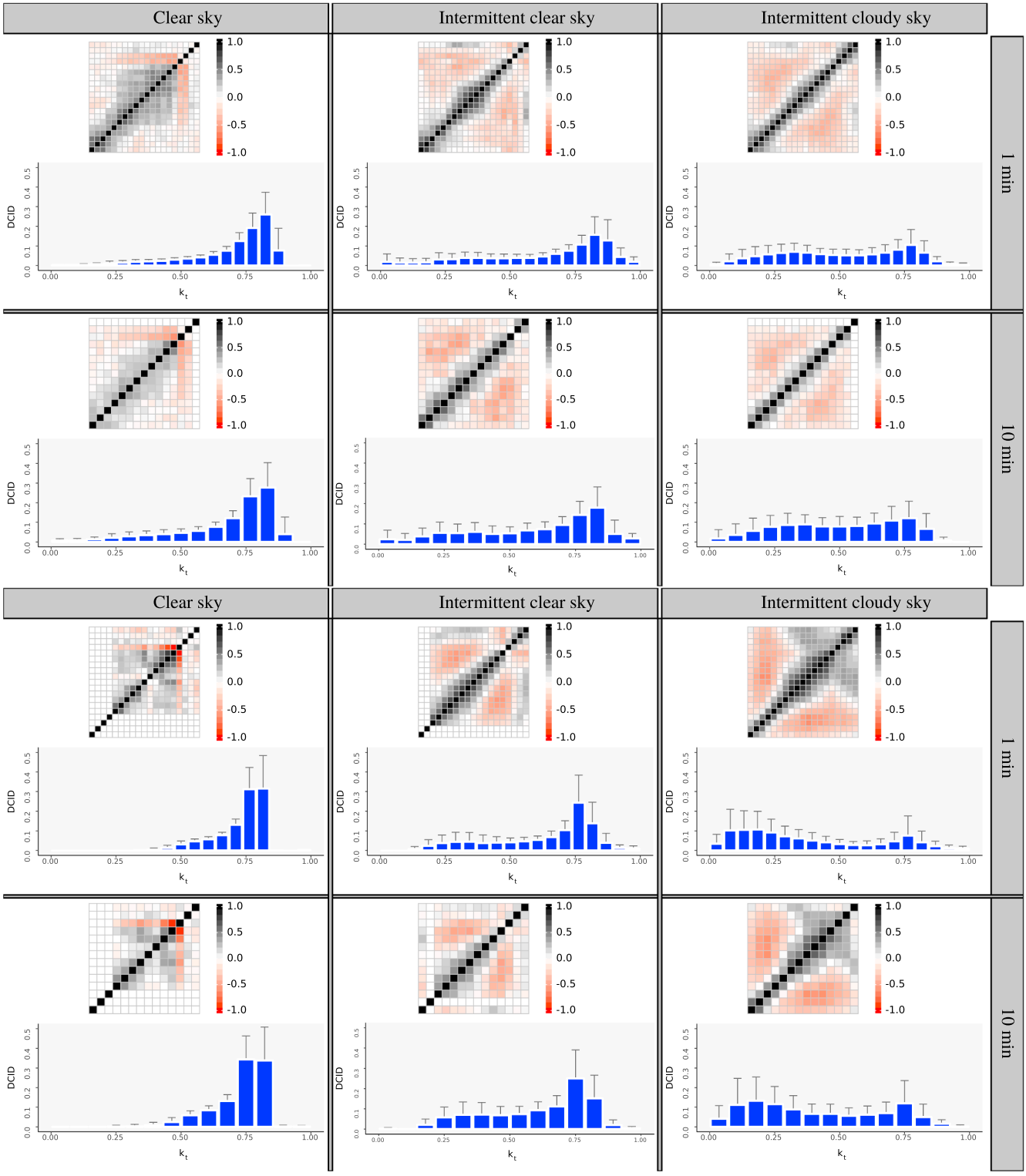

Fig. 3. Clustering results of Tamanrasset, Algeria (TAM) in the top plot and Alice Springs, Australia (ASP) in the bottom plot for each time step: blue histograms represents the mean daily clearness index distribution of the classes, heatmaps represent the correlation matrices of the Gaussian likelihoods. The direction of the increasing values in the heatmaps is from left to right and bottom to top. Error-bars represent the standard deviations at each bin for daily clearness index distributions in each class. DCID on the x-axis for the histograms stand for daily clearness index distribution. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

underlying structures in the data at different sampling rates. This is fundamentally achieved through the selection of the appropriate k_t daily distribution bin numbers. Starting with this point and taking into account that similar days should have similar statistical dynamic characteristics resulting from specific weather conditions, we infer the underlying transition matrix of each cluster after concatenating all its days and discretizing their values into a sufficient number of states. The transition matrix is calculated only to ensure that the order of the generated values take into consideration the typical dynamic of data values in each cluster. In fact, the inference of a transition matrix for

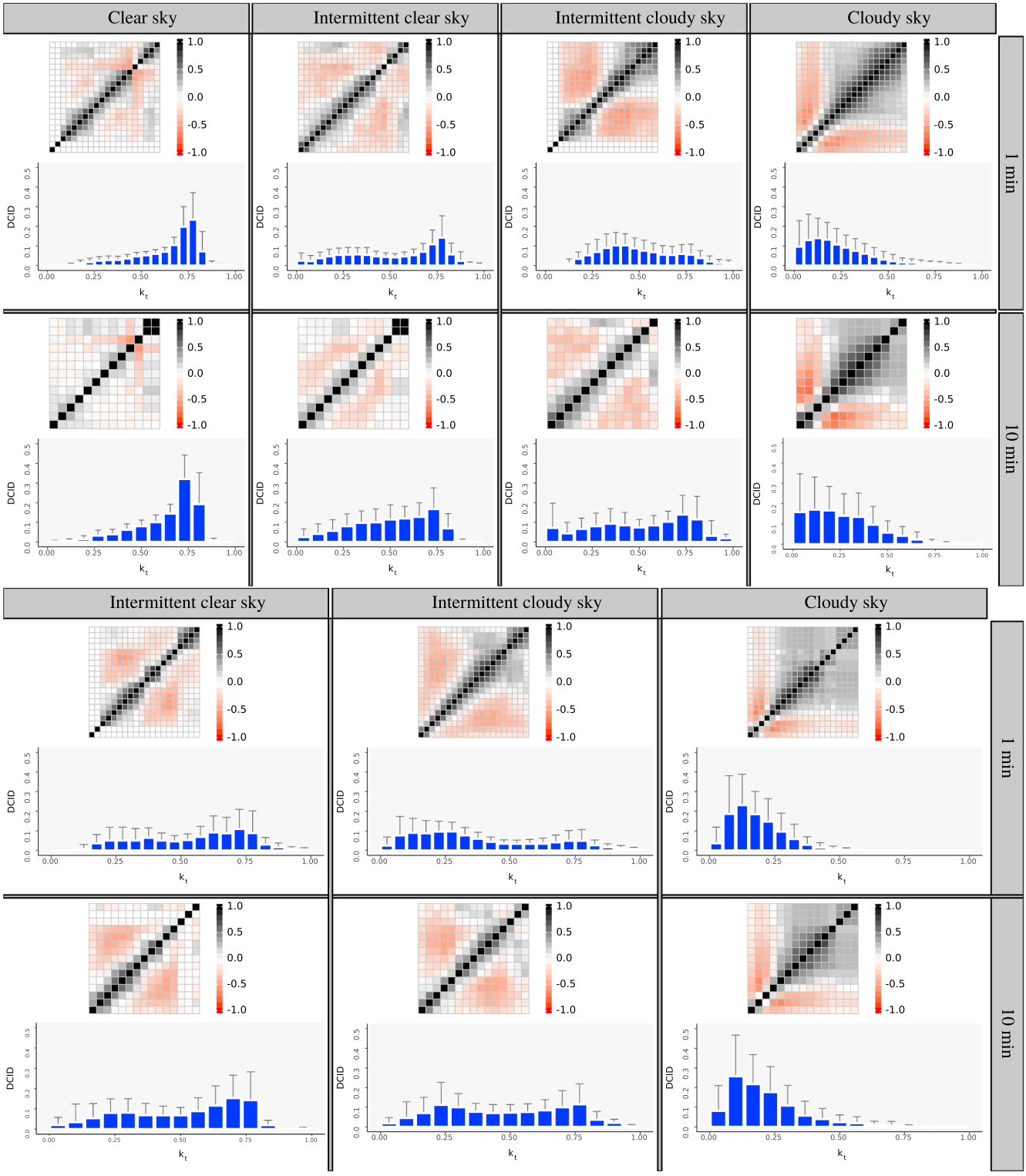

Fig. 4. Clustering results of Sioux Falls, USA (SXF) in the top plot and Toravere, Estonia (TOR) in the bottom plot for each time step: blue histograms represents the mean daily clearness index distribution of the classes, heatmaps represent the correlation matrices of the Gaussian likelihoods. The direction of the increasing values in the heatmaps is from left to right and bottom to top. Error-bars represent the standard deviations at each bin for daily clearness index distributions in each class. DCID on the x-axis for the histograms stand for daily clearness index distribution. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

each class guaranteed that the stationary distribution of the underlying Markov chain is the corresponding component of the class in the mixture distributions rendered by the DPGMM. Another feature of the methodology is that it can generate data based on available similar climate clustering. On the basis of the resulting mixture distribution of an available DPGMM classification under the same climatic conditions, we select the corresponding cluster for each day from the new database by maximizing the likelihood function of each component of the mixture. Afterwards, we carry out the synthetic generation procedure.

Algorithm 2.

Data: Database with a valid number of intraday observations to construct h_t
Result: 1-min GHI time-series
Summarize k_t as daily equal width histogram h_t;
if *Using similar climate* **then**
  Label the input data-set by assigning each day to the class that gives the maximum likelihood;
end
else
  Cluster the input data-set with DPGMM;
end
for *i in 1:number of clusters* **do**
  Concatenate and discretize the k_t time-series of the cluster i;
  Infer the underlying transition matrix TM_i;
end
Calculate sd-set: daily standard deviations of increments;
for *j in 1:number of days* **do**
  Assign to z_j the class label of the day j;
  for *k in 1:number of states in the day j* **do**
    Initialize the log-probability of the new sequence of states: $ps_{fin} = \inf$;
    Initialize by a random sequence of states seq;
    Calculate the increment inc_k;
    for *T in 1:number of iterations* **do**
      Generate a new sequence $nseq$ based on the corresponding TM_{z_j};
      if *nseq did not end with the state* $k+1$ **then**
        $ps_{fin} = ps_{fin}$;
        $seq = seq$;
      end
      else
        Calculate the probability ps_T of the sequence;
        if *(sd_k and inc_k are $>$ 70% of sd-set values) or ($sd_k > 95\%$ and $inc_k > 10\%$ of sd-set values);* **then**
          Consider $nseq$ with high sd: $ps_T = -sd_k$;
        end
        else
          Consider: $ps_T = ps_T * sd_k$;
        end
        if $ps_{T-1} > ps_T$ **then**
          $ps_{fin} = ps_T$;
          $seq = nseq$;
        end
      end
    end
  end
end

4. Experimental results and discussion

4.1. Similarity of the number of classes

In this section, we evaluate the class similarity observed from using the DPGMM with 1-min and 10-min temporal resolution irradiance time series. The 10-min time series is obtained as a direct averaging of the 1 min-time data.

Fig. 3 shows the clustering results of ASP and TAM data for 1-min and 10-min temporal resolution irradiance; Fig. 4 shows this for SXF and TOR data. These plots provide an intuitive geometric interpretation of the mean vector and the variance–covariance matrix of each class. The variance–covariance matrix is represented as follows: (1) error bars describe the spread of the data in the parallel directions to the axes of the h_t space (standard deviations); (2) a graphical display of the correlation matrix to represent the covariance of the i-th axis of the h_t space with the j-th one. In terms of probabilistic modeling, classes of solar irradiation conditions replace the DPGMM components. Thus, a class j is explained completely by its mean vector μ_j and its variance–covariance matrix Λ_j^{-1}.

It can be clearly noted from all plots that data for a given site have the same number of classes and share a high similarity for both the mean vector and the correlation graph at all time steps of the investigation. Furthermore, Fig. 5 demonstrates a reasonable similarity between class weights for all sites and time steps.

The clearness index characterizes the sky conditions of a particular place and time. A low clearness index value expresses a small quantity of the extraterrestrial horizontal solar irradiance reaching the earth's surface, generally representing a cloudy sky; a high clearness index value (typically around 0.8) implies that much of the global solar irradiance reaches the surface, representing a clear sky.

For ASP and TAM (BWh climate), three classes were found to represent the data: clear days, intermittent clear days and intermittent cloudy days; four classes represented SXF (Dfb climate) and three for TOR (Dfb climate). Clear days at TOR are often quite variable, this is most attributable to the highly turbid atmosphere and high albedo at the site. This results in the absence of a very clear class at TOR that exists at SXF.

The clear sky conditions corresponding to a monomodal mean distribution of k_t with a high occurrence of values around 0.7. The correlation graphs indicate that there is no relationship between the high values of k_t and the lower values in this class. In other words, as the frequency of the high values increases, the frequency of the lower values decreases, indicating a lower probability of fluctuation and a steady radiation during the day. This class of solar irradiance can be justified by a good atmospheric transmittance with few high, slow and thin clouds (cirrus, cirrocumulus).

The clear class and the intermittent clear class are very similar in mean distribution. This is because both classes essentially capture clear-sky periods for the majority of the time. From the correlation matrix, however, we found that the low and high values of the intermittent class are more correlated than their equivalents in the clear class. These types of distribution are regularly described as bi-modal (Smith et al., 2017) as there is a dominant peak representing those periods of clear sky (around $k_t = 0.7$–0.8) and a secondary peak that represents measurements under cloud or at low solar elevation.

Under intermittent cloudy sky conditions, the mean is a bimodal distribution, one mode around $k_t = 0.25$, and the other around $k_t = 0.7$. This class can be explained by a sunny regime mixed with a considerable number of clouds, which are often darker and composed primarily of water droplets (altocumulus, altostratus).

Through cloudy sky conditions, most of the solar irradiation is scattered or reflected due to several opaque clouds that are mainly composed of water droplets and which have a low dynamic level, covering almost the entire sky and combined with significant atmospheric turbidity (cumulus, stratocumulus).

4.2. Similarity of the class sequences

The main objective of class sequence analysis is to measure the similarity between sequences obtained at different data time step. Fig. 6 offers a view of the transversal distributions (TDs) (Gabadinho et al., 2011) of classes over days. TDs are transverse statistical characteristics computed at each day of the considered sequences. Each TD is provides the proportions of the resulting classes at the different time steps for each day. In the same vein, Fig. 6 shows the duration spent in each class through the length of each color segment (number of consecutive days in the same class).

To indicate how each sequence is organized, we use the transition rate between each pair of classes. It is calculated as the count of transitions between each two solar irradiation classes in the sequence

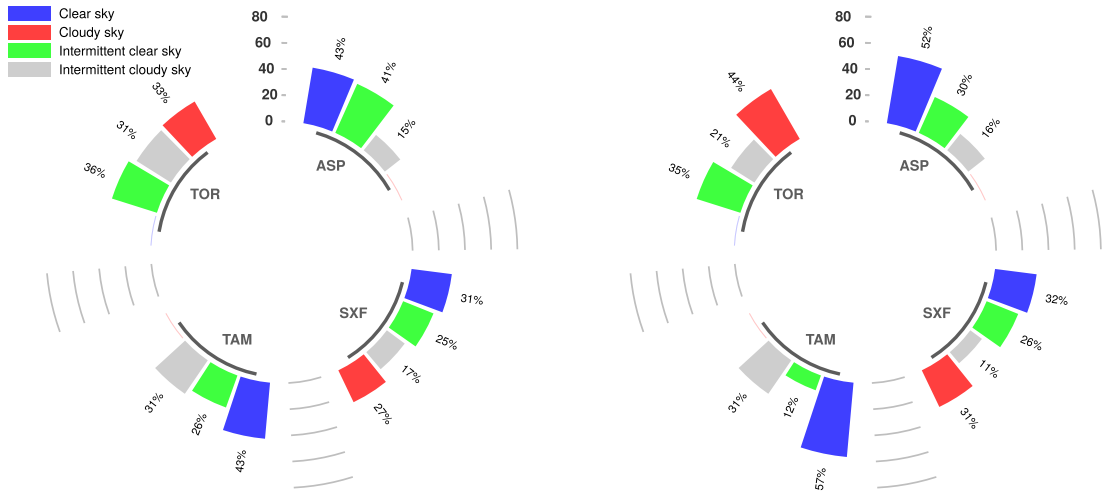

Fig. 5. Class weights for all sites for: (left) 1-min data and (right) 10-min data.

divided by the total number of transitions. Table 3 shows the transition rate matrix (TRM) for each time step (their elements are the transition rate between classes) in each site. It can be seen that TRMs are close to each other for a given site.

Occasionally, there is disagreement to which class a day belongs to depending on the averaging period. Whilst most of the time the class is the same between 1-min and 10-min logging intervals, they can also disagree and instead assign the day to an adjacent class (the next closest class, e.g., clear using 1-min data and intermittent clear using 10-min data). This observation is due to the similarity of the borders between adjacent classes, furthermore, the number of bins within the k_t distribution change at different averaging periods. For an effective visualization of the similarity of the class sequences, we count the frequency of states that are identically and non-identically classified between the resulting class sequences at different time averaging periods for each site. Fig. 7 shows these frequency of agreement between class assignment between the different averaging periods. A means the day is classified the same. B means that days are not classified the same and are instead assigned to an adjacent class. C means that the day is classified differently by two non-adjacent classes. All frequency

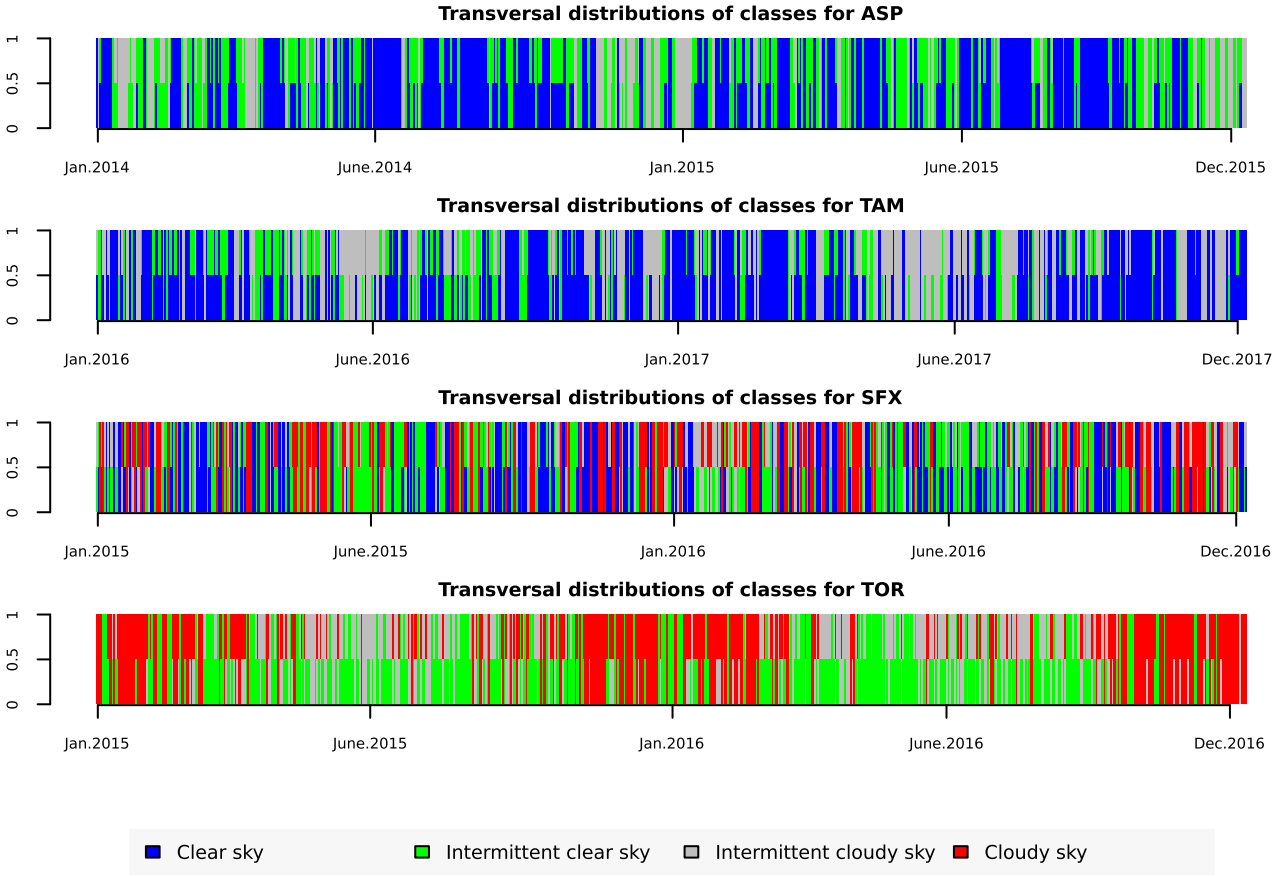

Fig. 6. Transverse distributions of classes for each meteorological station throughout the simulation period. Each day can be constructed of any of the identified classes belonging to that site.

Table 3
Transition rate matrices of all sites for each time step; C_1: Clear sky, C_2: Intermittent clear sky, C_3: Intermittent cloudy sky, C_4: Cloudy sky.

ASP:	1 min			10 min		
	→ C_1	→ C_2	→ C_3	→ C_1	→ C_2	→ C_3
C_1→	0.73	0.25	0.02	0.76	0.20	0.04
C_2→	0.26	0.58	0.16	0.37	0.40	0.23
C_3→	0.04	0.41	0.55	0.16	0.30	0.55

TAM:	1 min			10 min		
	→ C_1	→ C_2	→ C_3	→ C_1	→ C_2	→ C_3
C_1→	0.62	0.21	0.18	0.73	0.10	0.17
C_2→	0.34	0.42	0.24	0.46	0.30	0.24
C_3→	0.22	0.21	0.57	0.34	0.10	0.56

SXF:	1 min				10 min			
	→ C_1	→ C_2	→ C_3	→ C_4	→ C_1	→ C_2	→ C_3	→ C_4
C_1→	0.45	0.34	0.08	0.14	0.48	0.29	0.07	0.16
C_2→	0.22	0.49	0.09	0.20	0.27	0.38	0.11	0.24
C_3→	0.26	0.38	0.18	0.18	0.32	0.32	0.19	0.18
C_4→	0.16	0.35	0.08	0.41	0.20	0.27	0.09	0.45

TOR:	1 min			10 Minutes		
	→ C_2	→ C_3	→ C_4	→ C_2	→ C_3	→ C_4
C_2→	0.53	0.26	0.21	0.45	0.26	0.29
C_3→	0.34	0.51	0.15	0.43	0.30	0.27
C_4→	0.19	0.20	0.61	0.23	0.13	0.64

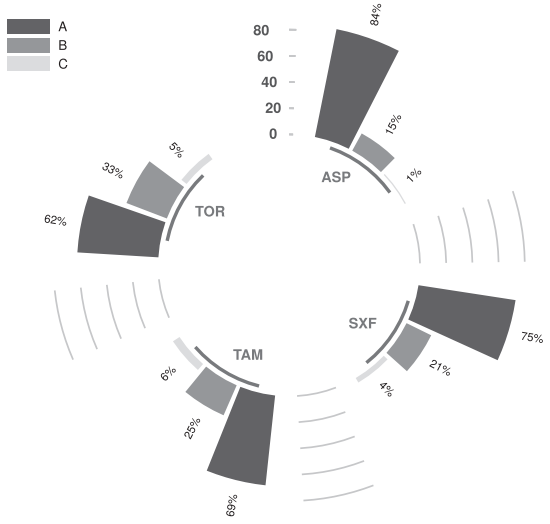

Fig. 7. Frequency of days assignment between the resulting class sequences at different time averaging periods for each site. A: frequency of days that are classified the same, B: frequency of days that are assigned to an adjacent class (the next closest class). C: frequency of days that are classified differently by two non-adjacent classes.

distributions have an exponential shape where the modes of their histograms correspond to a perfect similarity between the class sequences, reaching up to 84% for ASP, 69% for TAM, 75% for SXF and 62% for TOR. Situations when a class is assigned to a non-adjacent class occurs less than 6% of the time (<1% for TAM).

These results demonstrate that the DPGMM can perform reasonable recognition of solar irradiance states at different data averaging periods. This cluster analysis could help grid operators or large scale solar installers identify regions that are likely to experience regular ramping conditions and solar variability, or contrastingly it could identify those regions that are more likely to deliver smooth power generation.

5. Application: Generation of synthetic time series

5.1. Data generation

In this section, we demonstrate that we can provide a useful scheme for synthesizing solar irradiation time-series for 1-min time resolution based on the results from using DPGMM clustering. The irradiance time series are averaged to 10-min resolution for input into the DPGMM; other temporal averaging resolutions could be selected to construct daily distributions. Here, the choice of the 10-min interval allows the construction of relevant h_t for high-altitude sites. As the DPGMM is fully flexible to the input data, the method of representing 10-min data or similar intervals does not change the methodology, however, the representativeness of the resulting clusters has a strong dependency on the input data. From our experience with the data in this study, we find that a 2-year sample size and greater is suitable for using an averaging approach.

To validate our method, we use six quantifiers of variability in solar irradiance time-series as suggested in Blaga and Paulescu (2018). These quantifiers capture different facets of solar irradiance variability making the analysis complete. They are:

- The standard deviation of the increments (SDI)
- stability index (ST)
- integrated complementary cumulative distribution function (ICDF)
- sunshine stability number (SSN)
- fractal dimension (FD)

Furthermore, we also propose the use of the mean of the increments (MI), which sufficiently (if not completely) characterize the distribution of the increments.

SDI index is the most used. It captures the average deviation of increments around its mean. A low SDI indicates a smooth shape of daily solar irradiance curve, while a high SDI indicates a fluctuating solar irradiance curve. Also, we want to emphasize that the SDI for the

clearness index has a wrong formula in the Blaga and Paulescu (2018) paper. The suggested correction in this paper is as follow:

$$SDI\,(dk_t) = \sqrt{\frac{\sum_{i=1}^{N}(|dk_{t,i}| - m)^2}{N-1}}, \tag{8c}$$

where $m = \frac{\sum_{i=1}^{N} |\,dk_{t,i}|}{N}$ is the MI values. N is the number of data in the considered time-interval. The use of $N-1$ instead of N produces an unbiased estimator for the SDI index.

ST quantifies the cumulative increments in absolute value over a time-interval Δt larger than the logging-interval of the solar irradiance data. It is considered unstable if the increment over the chosen time-interval is larger than a threshold value. Following Tomson and Tamm (2006) we take the threshold for the increments in solar irradiance time-series to be 500 Wm^{-2} and Δt time-interval is 10-min. ICDF is obtained by integrating the complementary ECDF of the cumulative increments. It gives the sum of all possible cumulative increments that characterizes the day. SSN indicates the frequency of the Sun occurrence in the sky. It is calculated using a random time-dependent binary variable that uses direct normal irradiance (DNI) values and zenith angle, for more details see Blaga and Paulescu (2018). Lastly, FD of a solar irradiance time-series is a measure of how much it is jagged. It ranges between 1 and 2. FD of value 1 correspond to a straight line and a FD of value 2 correspond to a wiggly line that is totally fills up a plan. Interested readers can find more detailed information in Harrouni (2008).

The validity of our methodology is examined by statistically quantifying the sameness of the empirical cumulative distribution functions (CDF) of each quantifier calculated from the synthesized data and the corresponding 1-min measured time-series. It is expected that each simulated CDF match their respective measured CDF. In this work we use two metrics to measure the goodness-of-fit between CDFs; the overlap coefficient (OVC), and Kullback–Leibler divergence (KLD). OVC is a measure of the similarity between two probability functions by measuring the amount of overlap between them. It ranges between 0 and 1. Higher values correspond to a better fit and smaller values correspond to two different data-sets. KLD is a measure of the number of bits of information lost when approximating one distribution by the other. It ranges between 0 and 1 and in contrast to the OVC, a KLD of 0 indicates that the two distributions are identical, and 1 indicates that are totally different.

Fig. 10 shows an example of two different generated days. More examples can be found in Appendix B. Fig. 8 shows the plots of the quantifier CDFs of observed and synthetically generated GHI. All CDFs are correlated. Table 4 shows the OVC and the KLD values obtained for each selected location. These values confirm the good fit between the measured and generated GHI series. The OVC is not less than 0.75 all the time and the KLD does not exceed 0.1, for all sites and quantifiers. To test the validity of our methodology to generate data at a different location than the training data, we apply testing to sites based on similar climate conditions. Each two similar climate databases are compared by the same manner described above. From Table 2 we compare ASP to TAM, SFX to TOR. Fig. 9 display the CDFs of each quantifier over the generated and measured data for each pair of locations. The CDFs are correlated. Table 5 confirms this correlation by showing high values for the OVC index and lower values for the KLD measure. This confirms that there is a good agreement when applying the methodology in a synthetic irradiance generation application. With only the training data from a few locations, it is suggested that the methodology can be applied to any other similar climate with reasonable results. This is very noticeable between ASP and TAM ground stations whereby the training and testing datasets are 14,783 km separated on different continents and hemispheres. We see fantastic reproducibility across all metrics. This can be also observed between SFX and TOR, two ground stations located also on two other different continents. This very much implies global applicability with an appropriate database of climate types to train on. The implications of this are that, for a given location, high-temporal resolution synthetic irradiance data can be generated so long as data from a similar climate exists, which it invariably does on account of the extensive collections from the BSRN.

The predictive accuracy in terms of normalized root mean square error (nRMSE) and normalized mean bias error (nMBE) are depicted in Table 6. It is observed that in terms of nRMSE, DPGMM shows a good similarity in the GHI means for both daily and monthly time scales with all values <0.04; this is also true for both local generation and similar climate generation indicated at the bottom half of the table where all nRMSE are <0.04. By showing the nMBE results, we conclude that the application of DPGMM generates an unbiased synthetic GHI time series with all nMBE within ± 0.02 and ± 0.04 for similar climates.

5.2. Discussion of model weaknesses

In this section, we discuss our perceived disadvantages to the proposed method. Firstly, the use case of such data analysis has a prerequisite for input data. Other synthetic irradiance generators (e.g. Bright et al. (2017) and Munkhammar and Widén (2018)) do not require irradiance as an input once certain training distributions have been derived, however, their derivations require those high resolution data sources which are not widely available. That said, certain synthetic irradiance generators (e.g. Ngoko et al. (2014)) require 1-min data prior to the methodology being applied. With our methodology, the requirement is time series of 10-min or lower temporal resolution data, which is readily available from satellite imagery as provided by providers such as SolarGIS, Solcast, SolarAnywhere etc. Geographically speaking, this makes our methodology more flexible in many respects despite the dependency on training data. This does raise the question as to what can be fully considered synthetic data and which should be considered downscaling? Our methodology is perhaps a hybrid, as it can be used for purely synthetic time series without the guidance of the input data, however, it can also very successfully attempt to downscale the data by gap filling with 1-min resolution that has the statistical properties of real data, even if it is not expected to match. Grantham et al. (2017) and Bright (2019) proposed all the reasons why this type of methodology is useful, however, we have avoided the need for storing a significant database of sequences in order to downscale. Our classification technique avoids the clear pitfalls of their sequence-based methodology whereby ramps were facilitated during actually clear periods because the mean hourly k_t was maintained even under highly variable skies.

Due to our advanced classification component, the number of classes for every site can be specifically tailored; however, there is an exception. When using the DPGMM with a small sample size (e.g., only 1 year), the total number of classes that physically exist at the site may not have been appropriately captured and so the site would not be properly represented. This is easily explained as the prior probabilities for a new class to exist are not satisfied with the reduced sample. As an example, if a site has four observable classes of 50 clear, 50 partially cloudy, 50 cloudy and only 2 intermittently clear days, the prior probabilities assigned would not enable this latter class to exist due to the reduced likelihood of it being an independent class. Hence, it is important to have a significantly sized and representative sample. This is a weakness in the model in that there must be a large enough sample size is required for optimal performance. In practise, we would use—and strongly recommend—sources of coarser resolution data which tend to have a more significant history. This is entirely suitable as most satellite-derived bankable solar irradiance databases extend to significant historical duration.

The recent findings of Bright (2019) that sites matched by climate did not provide the best synthetic generation than other sites that inherently contained a more appropriate statistical similarity. It was suggested that variability was the crucial component. Furthemore, Schwarz et al. (2018) mapped the whole world in regards to spatial

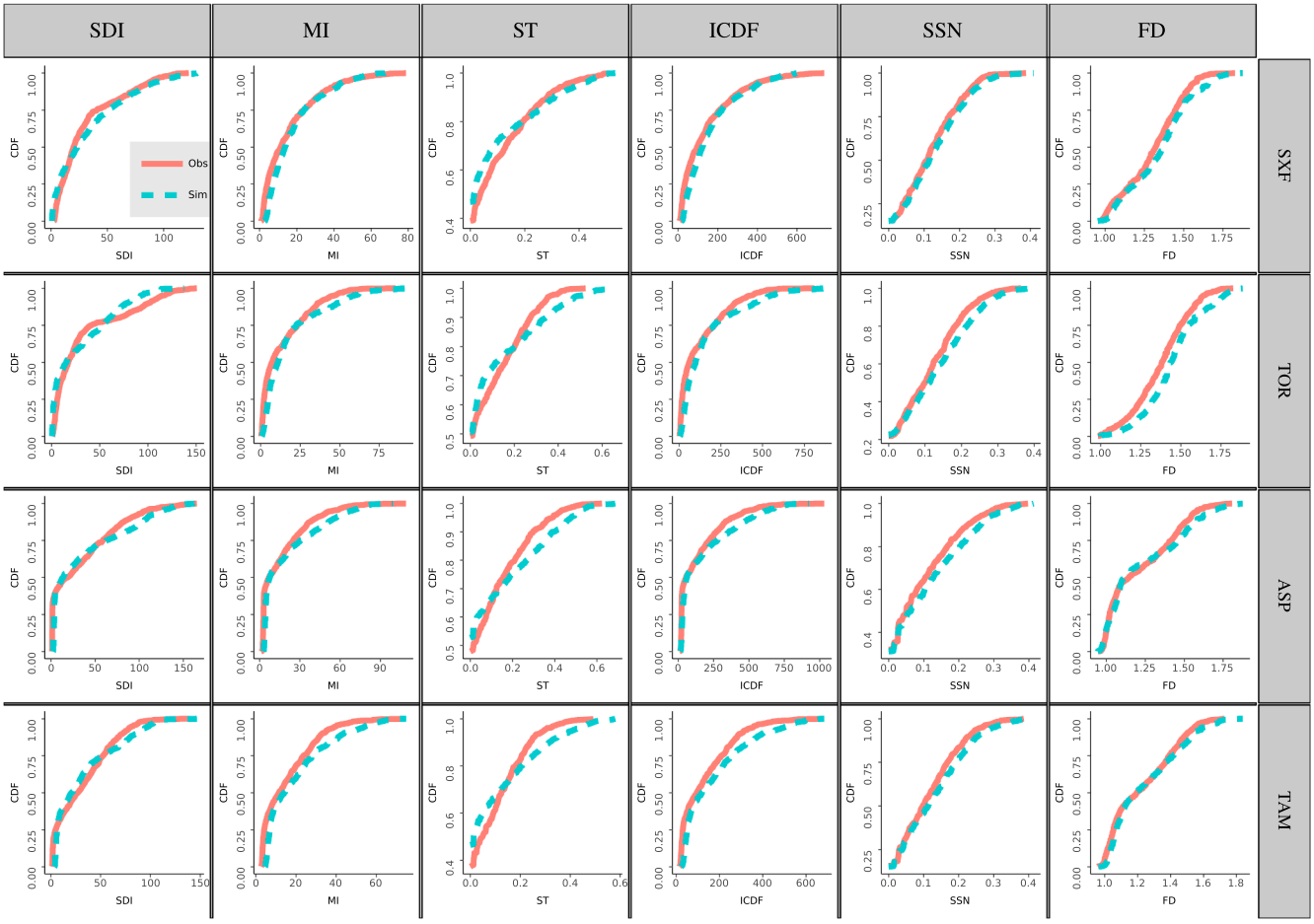

Fig. 8. Comparison between observation data CDFs (red solid line) and synthetic data CDFs (blue dashed line) using six indicators of solar irradiation variability. Each row represents one of the ground measurement sites used in the study and each column represents one of the used solar irradiance quantifiers. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
The overlapping coefficient (OVC) and the Kullback–Leibler divergence (KLD) metrics calculated between each distribution pair represented in Fig. 8. Each row represents one of the ground measurement sites used in the study and each column represents one of the used solar irradiance quantifiers.

	SDI	MI	ST	ICDF	SSN	FD
			OVC			
SXF	0.86	0.92	0.85	0.90	0.95	0.91
TOR	0.75	0.80	0.80	0.80	0.85	0.84
ASP	0.80	0.81	0.80	0.81	0.90	0.87
TAM	0.80	0.82	0.75	0.82	0.90	0.88
			KLD			
SXF	0.03	0.03	0.09	0.07	0.05	0.05
TOR	0.10	0.06	0.10	0.04	0.04	0.04
ASP	0.06	0.09	0.02	0.10	0.01	0.01
TAM	0.06	0.07	0.05	0.07	0.01	0.01

variability correlation and presented fascinating maps of variability similarity. We expect that our methodology would be able to appropriately represent the world so long as the classes at each site are appropriately considered, perhaps as guided by the Schwarz et al. (2018). At present, we train the DPGMM on similar climates, and so our application is potentially flawed by selecting training data that is not necessarily the most representative, Even so, as our methodology is a synthetic-downscaling blend, the more granular data helps guide the variability before a 1-min time series is synthetically generated. Another weakness is the inability to perform with hourly data that is far more readily available. The training of the DPGMM requires a distinct distribution from each day's data in order to establish a class, using hourly data does not facilitate a clear distribution.

The use of conjugate priors in the computation and analysis is not as appropriate as using non-conjugate priors, as the latter would match the reality with a significantly higher accuracy. For computation, it is far easier to manage conjugate priors. Already the computation time for DPGMM is not as streamlined as alternative methodologies, thus presenting a further weakness of the proposed methodology. We do, however, believe that the resultant irradiance time series are more appropriate than alternatives.

Perhaps the most significant weakness in the field of synthetic irradiance generation is the inability to model the spatial dimension. The science of synthetic irradiance generation is quickly moving towards simulations of distributed PV installations which require spatio-temporal time series. Whilst the current presented version of the DPGMM does not contain the spatial correlation, it could be included by including spatial information within the clustering approach by pairing another layer of prior distributions, however, this is not provided or explored at present.

6. Conclusion and future perspective

In this paper, we propose a consistent way for automatic classification of k_t distributions with a flexible and robust specification of the likelihood and prior distributions, namely Dirichlet process Gaussian mixture model (DPGMM). It is a nonparametric Bayesian model where the complexity level and size are not specified in advance and may

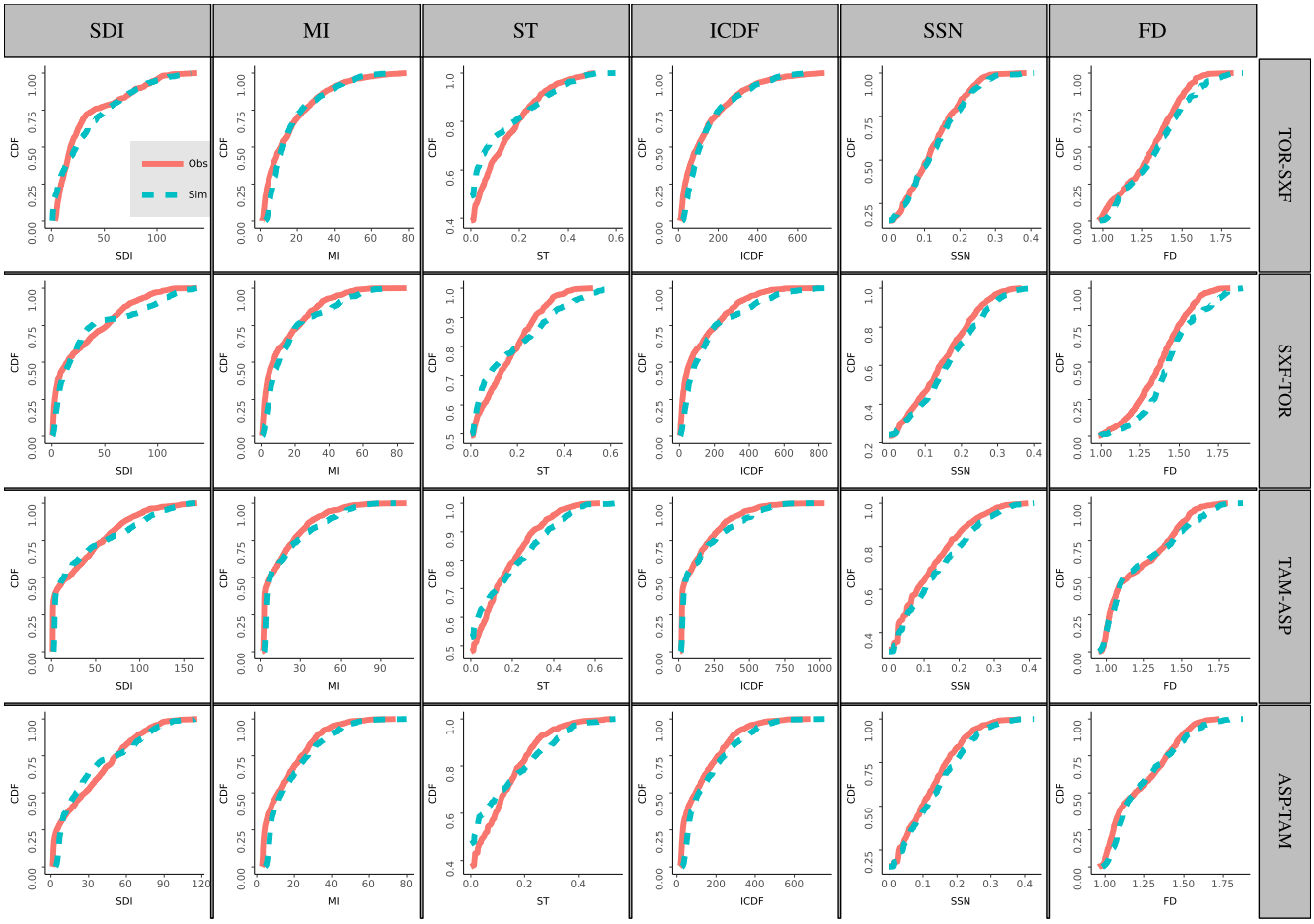

Fig. 9. Data generation based on similar climate classification: observation data (red solid line), synthetic data (blue dashed line). The row labels are in the format (training–testing), e.g. TAM-ASP means training on Tamanrasset and testing on Alice Springs. Each row represents one of the ground measurement sites used in the study and each column represents one of the used solar irradiance quantifiers. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 5
The overlapping coefficient (OVC) and the Kullback–Leibler divergence (KLD) metrics calculated between each distribution pair represented in Fig. 9. Each row represents one of the ground measurement sites used in the study and each column represents one of the used solar irradiance quantifiers.

	SDI	MI	ST	ICDF	SSN	FD
			OVC			
TOR-SXF	0.82	0.89	0.81	0.90	0.94	0.87
SXF-TOR	0.80	0.86	0.80	0.87	0.93	0.80
TAM-ASP	0.80	0.84	0.83	0.85	0.90	0.87
ASP-TAM	0.80	0.87	0.80	0.88	0.90	0.85
			KLD			
TOR-SXF	0.09	0.10	0.10	0.01	0.03	0.03
SXF-TOR	0.05	0.09	0.06	0.09	0.07	0.07
TAM-ASP	0.04	0.04	0.02	0.05	0.01	0.01
ASP-TAM	0.07	0.02	0.04	0.02	0.03	0.03

Table 6
Table of normalized root mean squared errors (nRMSE) and the normalized mean bias errors (nMBE) between observed and generated daily and monthly means.

	Daily mean		Monthly mean	
	nRMSE	nMBE	nRMSE	nMBE
		Local generation		
SXF	0.02	−0.01	0.02	−0.02
TOR	0.02	−0.01	0.02	−0.02
ASP	0.02	−0.01	0.03	−0.03
TAM	0.02	−0.01	0.04	−0.04
		Similar climate generation		
TOR-SXF	0.02	−0.01	0.02	−0.02
SFX-TOR	0.02	−0.01	0.03	−0.02
TAM-ASP	0.02	−0.01	0.03	−0.03
ASP-TAM	0.02	−0.01	0.04	−0.03

increase as new data come in. The originality of this framework is that it can deal with a variety of data sampling rates, and there is no need to specify any parametric assumption or to restrict the number of solar radiation classes. This clustering method is successfully applied in the generation of temporal synthetic 1-min solar irradiance time-series whereby many of the issues of previous synthetic generation methodologies are addressed.

The experimental databases are recorded at four solar irradiance monitoring stations around the world. To assess the consistency of the DPGMM, we have summarized the daily clearness indexes k_t as daily equal width histograms for different databases at different temporal resolutions. A good agreement between the resulting classes and their sequences is obtained.

Based on the resulting class distributions and the transition rules encoded by transition rate matrices, combined with a Markov model, we generated synthetic 1-min GHI time-series with minimal error compared to measured data. The global applicability of this methodology was demonstrated by training the methodology on one site and testing at a different location that shared similar climatic properties. We

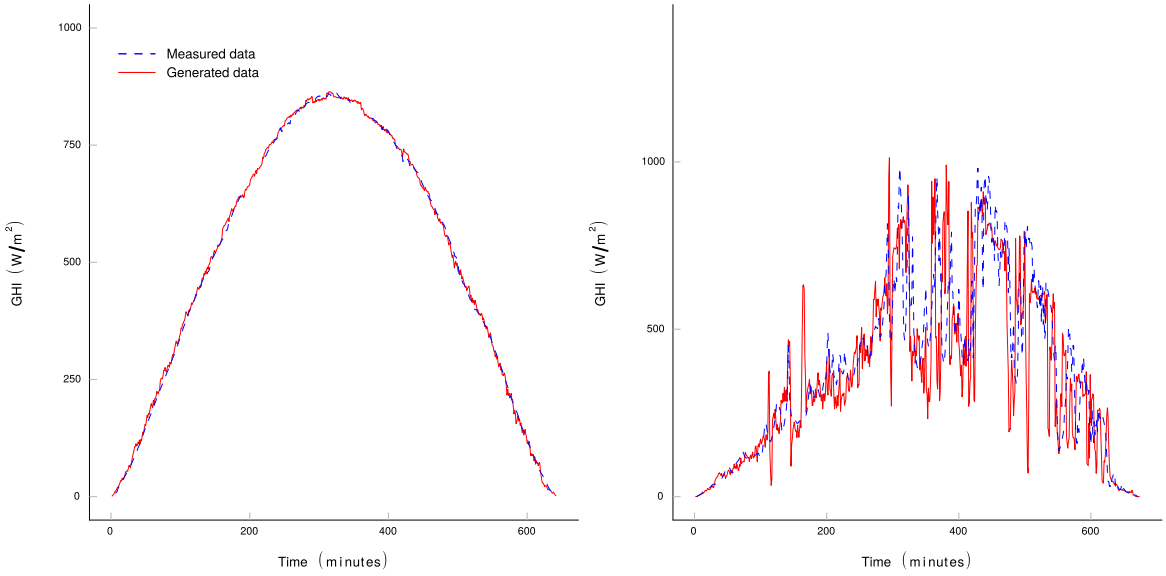

Fig. 10. An example of the generated synthetic GHI time series from the DPGMM alongside and observation data measured over two different days; left: clear sky day, right: overcast day. More extensive examples can be found in Appendix B.

observed very good performance at a globally applicable synthetic irradiance generation through our advanced clustering approach. For both the same location and when applied to similar climate-based generation, the method presents a Kullback–Leibler coefficient between measured and generated distributions of $\leqslant 0.1$ and an overlapping coefficient of $\geqslant 0.75$, for all sites. In addition, an nRMSE $\leqslant 0.04$ and nMBE $< \pm\, 0.04$ between generated and measured means for both daily and monthly scales in all cases are showed by using DPGMM.

Finally, we believe that apart from looking for separate locations, future research should look for the hierarchization of the DPGMM by sharing the resulting clusters between several locations in a territory. This would allow DP to be used as building blocks within a hierarchical model. Thus, we can demonstrate a new way to create new geographic information system data and solar maps at very high temporal resolution. Further complexities could be added to the specific application of the DPGMM to GHI, for example, the inclusion of annual variability and seasonality could be introduced.

To encourage other researchers to re-use our algorithm, this methodology is freely available as an R-package downloadable from https://github.com/frimane/SolarClusGnr.

Appendix A. Likelihoods

The multivariate Gaussian distribution (MGD) of a D-dimensional random vector X is:

$$\mathcal{N}(X) = \frac{1}{\sqrt{(2\pi)^D |\mathbf{\Lambda}|^{-1}}} \exp\left(-\frac{1}{2}(X-\mu)^T \mathbf{\Lambda}(X-\mu)\right), \tag{A.1}$$

where μ is the mean vector and Λ is the precision matrix.

The Wishart distribution of a $D \times D$ positive definite matrix of random variables Λ is:

$$\mathcal{N}(\Lambda) = \frac{1}{2^{\frac{\nu D}{2}} |W|^{\frac{\nu}{2}} \Gamma_D(\frac{\nu}{2})} |\Lambda|^{\frac{\nu-D-1}{2}} e^{-\frac{1}{2}\mathrm{tr}(W^{-1}\Lambda)} \tag{A.2}$$

where ν is the degree of freedom and W a fixed positive definite matrix of size $D \times D$. Γ_D is the multivariate gamma function.

The gamma distribution in the shape-scale (α-β) characterization is:

$$\mathcal{G}(x|\alpha, \beta) = \frac{x^{\alpha-1} e^{-x/\beta}}{\beta^{-\alpha}\Gamma(\alpha)} \quad \text{for } x > 0 \text{ and } \alpha, \beta > 0, \tag{A.3}$$

where $\Gamma(\alpha)$ is the gamma function.

The Dirichlet distribution of p, an element of the $D-1$ simplex, has the probability density function:

$$Dir(p) = \frac{\Gamma(\sum_{i=1}^{D} \alpha_i)}{\prod_{i=1}^{D} \Gamma(\alpha_i)} \prod_{i=1}^{D} p_i^{\alpha_i - 1}, \tag{A.4}$$

with α is the concentration parameter.

The GEM construction (stick-breaking process) is:

$$\pi_j = \beta_j \cdot \prod_{i=1}^{j-1} (1-\beta_i), \tag{A.5}$$

where $\beta_j \sim Beta(1, \alpha)$ for j from 1 to infinity. Not that Dirichlet distribution is the multivariate generalization of the *Beta* distribution.

Appendix B. Examples of days

See Fig. B.11.

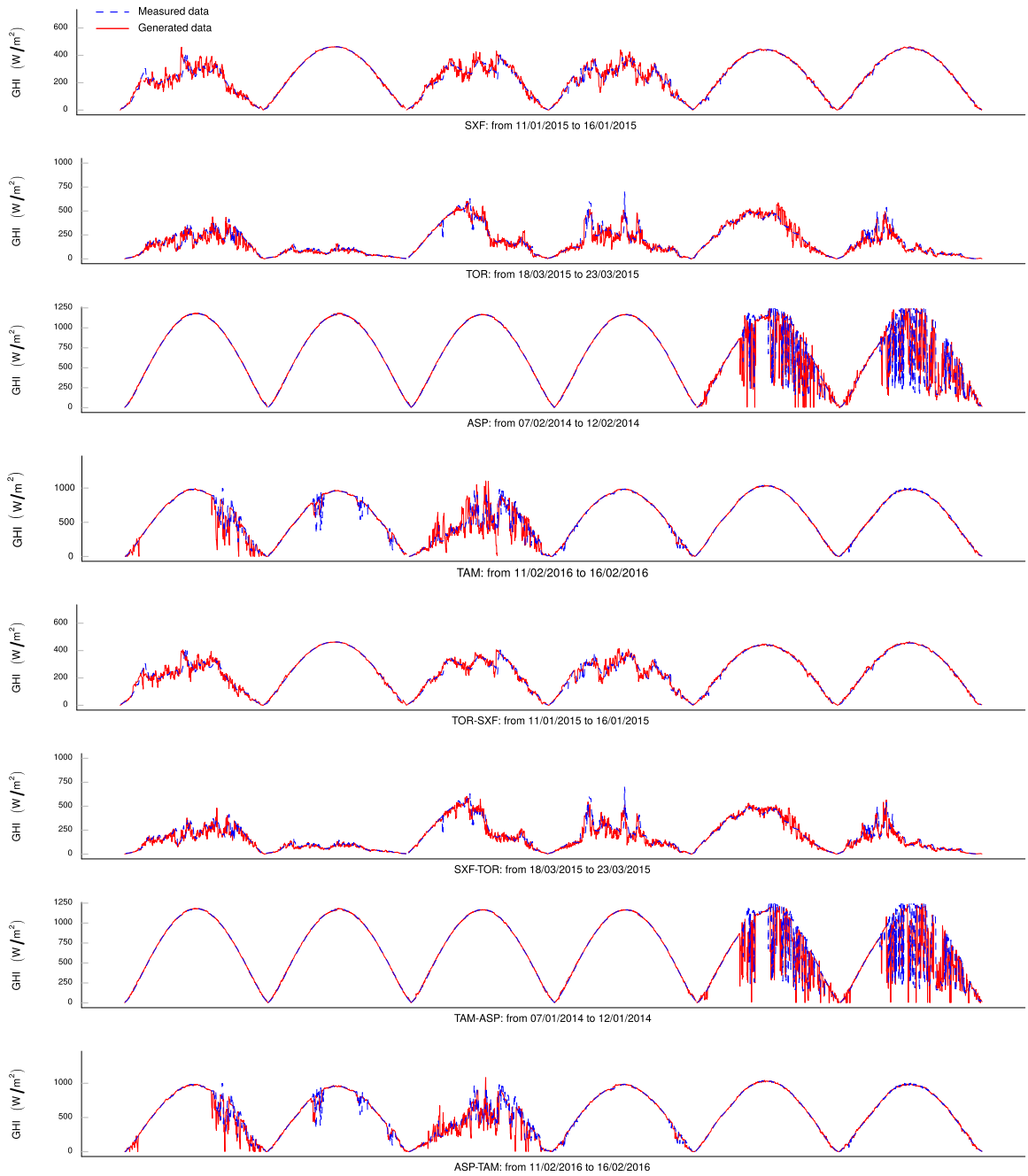

Fig. B.11. The days selected are continuous and were randomly selected purely for example purposes where there was a wide variety of cloud conditions, the range of days and site name are indicated in the x-axis label. The actual observed measurements are indicated in dashed blue lines whereas the generated synthetic data from the DPGMM are shown in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Appendix C. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.solener.2019.02.052.

References

Blaga, R., Paulescu, M., 2018. Quantifiers for the solar irradiance variability: A new perspective. Sol. Energy 174, 606–616. https://doi.org/10.1016/j.solener.2018.09.034. <http://www.sciencedirect.com/science/article/pii/S0038092X1830906X>.

Bright, J., Smith, C., Taylor, P., Crook, R., 2015. Stochastic generation of synthetic minutely irradiance time series derived from mean hourly weather observation data. Sol. Energy 115, 229–242. https://doi.org/10.1016/j.solener.2015.02.032. <http://www.sciencedirect.com/science/article/pii/S0038092X15001024>.

Bright, J.M., 2019. The impact of globally diverse GHI training data: evaluation through application of a simple markov chain downscaling methodology. Renew. Sustain.

Energy 11.
Bright, J.M., Babacan, O., Kleissl, J., Taylor, P.G., Crook, R., 2017. A synthetic, spatially decorrelating solar irradiance generator and application to a LV grid model with high PV penetration. Sol. Energy 147, 83–98. https://doi.org/10.1016/j.solener.2017.03.018. <http://www.sciencedirect.com/science/article/pii/S0038092X17301780>.
Bright, J.M., Killinger, S., Lingfors, D., Engerer, N.A., 2018. Improved satellite-derived PV power nowcasting using real-time power data from reference PV systems. Sol. Energy 168, 118–139.
Brock, G., Pihur, V., Datta, S., Datta, S., 2008. clvalid: An r package for cluster validation. J. Stat. Softw. 25, 1–22. https://doi.org/10.18637/jss.v025.i04. <ttps://www.jstatsoft.org/v025/i04>.
Engerer, N., Bright, J., Killinger, S., 2017. Himawari-8 enabled real-time distributed PV simulations for distribution networks. In: Proceedings of IEEE PVSC, pp. 25–30.<https://www.researchgate.net/publication/317845328_Himawari-8_enabled_real-time_distributed_PV_simulations_for_distribution_networks>.
Ferguson, T.S., 1973. A bayesian analysis of some nonparametric problems. Ann. Stat. 1, 209–230.<http://www.jstor.org/stable/2958008>.
Frimane, Â., Aggour, M., Ouhammou, B., Bahmad, L., 2018. A dirichlet-multinomial mixture model-based approach for daily solar radiation classification. Sol. Energy 171, 31–39. https://doi.org/10.1016/j.solener.2018.06.059. <http://www.sciencedirect.com/science/article/pii/S0038092X1830611X>.
Gabadinho, A., Ritschard, G., M?ller, N., Studer, M., 2011. Analyzing and visualizing state sequences in r with traminer. J. Stat. Softw. 40, 1–37. https://doi.org/10.18637/jss.v040.i04. <https://www.jstatsoft.org/v040/i04>.
García, S., Luengo, J., Sáez, J.A., López, V., Herrera, F., 2013. A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. IEEE Trans. Knowledge Data Eng. 25, 734–750. https://doi.org/10.1109/TKDE.2012.35.
Gershman, S.J., Blei, D.M., 2012. A tutorial on bayesian nonparametric models. J. Math. Psychol. 56, 1–12. https://doi.org/10.1016/j.jmp.2011.08.004. <http://www.sciencedirect.com/science/article/pii/S002224961100071X>.
Ghayekhloo, M., Ghofrani, M., Menhaj, M., Azimi, R., 2015. A novel clustering approach for short-term solar radiation forecasting. Sol. Energy 122, 1371–1383. https://doi.org/10.1016/j.solener.2015.10.053. <http://www.sciencedirect.com/science/article/pii/S0038092X1500609X>.
Görür, D., Edward Rasmussen, C., 2010. Dirichlet process gaussian mixture models: Choice of the base distribution. J. Comput. Sci. Technol. 25, 653–664.
Grantham, A., Pudney, P., Ward, L., Belusko, M., Boland, J., 2017. Generating synthetic five-minute solar irradiance values from hourly observations. Sol. Energy 147, 209–221. https://doi.org/10.1016/j.solener.2017.03.026. <http://www.sciencedirect.com/science/article/pii/S0038092X1730186X>.
Gueymard, C.A., 2018. A reevaluation of the solar constant based on a 42-year total solar irradiance time series and a reconciliation of spaceborne observations. Sol. Energy 168, 2–9. https://doi.org/10.1016/j.solener.2018.04.001. advances in Solar Resource Assessment and Forecasting. <http://www.sciencedirect.com/science/article/pii/S0038092X18303463>.
Gueymard, C.A., Myers, D.R., 2009. Evaluation of conventional and high-performance routine solar radiation measurements for improved solar resource, climatological trends, and radiative modeling. Sol. Energy 83, 171–185.
Harrouni, S., 2008. Modeling Solar Radiation at the Earth Surface. Springer, Berlin, Heidelberg chapter Fractal classification of typical meteorological days from global solar irradiance: Application to five sites of different climates. pp. 29–54.
Inman, H.F., Jr, E.L.B., 1989. The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. Commun. Stat. - Theory Methods 18, 3851–3874. https://doi.org/10.1080/03610928908830127. arXiv:https://doi.org/10.1080/03610928908830127.
Kahle, D., Wickham, H., 2013. ggmap: Spatial visualization with ggplot2. The R J. 5, 144–161.<http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf>.
Kang, B.O., Tam, K.S., 2013. A new characterization and classification method for daily sky conditions based on ground-based solar irradiance measurement data. Sol. Energy 94, 102–118. https://doi.org/10.1016/j.solener.2013.04.007. <http://www.sciencedirect.com/science/article/pii/S0038092X13001400>.
Kullback, S., Leibler, R.A., 1951. On information and sufficiency. Ann. Math. Statist. 22, 79–86. https://doi.org/10.1214/aoms/1177729694.
Larrañeta, M., Fernandez-Peruchena, C., Silva-Pérez, M., Lillo-Bravo, I., 2018. Methodology to synthetically downscale DNI time series from 1-h to 1-min temporal resolution with geographic flexibility. Sol. Energy 162, 573–584. https://doi.org/10.1016/j.solener.2018.01.064. <http://www.sciencedirect.com/science/article/pii/S0038092X18300859>.
Lave, M., Reno, M.J., Broderick, R.J., 2015. Characterizing local high-frequency solar variability and its impact to distribution studies. Sol. Energy 118, 327–337. https://doi.org/10.1016/j.solener.2015.05.028. <http://www.sciencedirect.com/science/article/pii/S0038092X15002881>.
Lee, Y., Wahba, G., Ackerman, S.A., 2004. Cloud classification of satellite radiance data by multicategory support vector machines. J. Atmosp. Oceanic Technol. 21, 159–169.
Long, C., Shi, Y., 2006. The QCRad value added product: Surface radiation measurement quality control testing, including climatology configurable limits. Atmospheric Radiation Measurement Program Technical Report.
Mateo, C., Cossent, R., Gómez, T., Prettico, G., Frías, P., Fulli, G., Meletiou, A., Postigo, F., 2018. Impact of solar PV self-consumption policies on distribution networks and regulatory implications. Sol. Energy 176, 62–72. https://doi.org/10.1016/j.solener.2018.10.015. <http://www.sciencedirect.com/science/article/pii/S0038092X18309940>.
Michalsky, J., Dutton, E., Rubes, M., Nelson, D., Stoffel, T., Wesley, M., Splitt, M., DeLuisi, J., 1999. Optimal measurement of surface shortwave irradiance using current instrumentation. J. Atmosph. Oceanic Technol. 16, 55–69.
Moreno-Tejera, S., Silva-Pérez, M., Ramírez-Santigosa, L., Lillo-Bravo, I., 2017. Classification of days according to DNI profiles using clustering techniques. Sol. Energy 146, 319–333. https://doi.org/10.1016/j.solener.2017.02.031. <http://www.sciencedirect.com/science/article/pii/S0038092X1730124X>.
Munkhammar, J., Widén, J., 2016. Correlation modeling of instantaneous solar irradiance with applications to solar engineering. Sol. Energy 133, 14–23. https://doi.org/10.1016/j.solener.2016.03.052. <http://www.sciencedirect.com/science/article/pii/S0038092X16300056>.
Munkhammar, J., Widén, J., 2018. An n-state markov-chain mixture distribution model of the clear-sky index. Sol. Energy 173, 487–495. https://doi.org/10.1016/j.solener.2018.07.056. <http://www.sciencedirect.com/science/article/pii/S0038092X18307205>.
Munkhammar, J., Widén, J., 2019. A spatiotemporal markov-chain mixture distribution model of the clear-sky index. Sol. Energy 179, 398–409. https://doi.org/10.1016/j.solener.2018.12.064. <http://www.sciencedirect.com/science/article/pii/S0038092X18312611>.
Munkhammar, J., Widén, J., Hinkelman, L.M., 2017. A copula method for simulating correlated instantaneous solar irradiance in spatial networks. Sol. Energy 143, 10–21. https://doi.org/10.1016/j.solener.2016.12.022. <http://www.sciencedirect.com/science/article/pii/S0038092X16306168>.
Munshi, A.A., Mohamed, Y.A.R., 2016. Photovoltaic power pattern clustering based on conventional and swarm clustering methods. Sol. Energy 124, 39–56. https://doi.org/10.1016/j.solener.2015.11.010. <http://www.sciencedirect.com/science/article/pii/S0038092X15006167>.
Neal, R.M., 2000. Markov chain sampling methods for Dirichlet process mixture models. J. Comput. Graphical Stat. 9, 249–265. https://doi.org/10.1080/10618600.2000.10474879. <https://www.tandfonline.com/doi/abs/10.1080/10618600.2000.10474879>.
Ngoko, B., Sugihara, H., Funaki, T., 2014. Synthetic generation of high temporal resolution solar radiation data using markov models. Sol. Energy 103, 160–170. https://doi.org/10.1016/j.solener.2014.02.026. <http://www.sciencedirect.com/science/article/pii/S0038092X14001042>.
Peel, M.C., Finlayson, B.L., McMahon, T.A., 2007. Updated world map of the köppen-geiger climate classification. Hydrol. Earth Syst. Sci. 11, 1633–1644. https://doi.org/10.5194/hess-11-1633-2007. <https://www.hydrol-earth-syst-sci.net/11/1633/2007/>.
Perez, R., Ineichen, P., Seals, R., Zelenka, A., 1990. Making full use of the clearness index for parameterizing hourly insolation conditions. Sol. Energy 45, 111–114. https://doi.org/10.1016/0038-092X(90)90036-C. <http://www.sciencedirect.com/science/article/pii/0038092X9090036C>.
Pérez-Ortiz, M., Jiménez-Fernández, S., Gutiérrez, P.A., Alexandre, E., Hervás-Martínez, C., Salcedo-Sanz, S., 2016. A review of classification problems and algorithms in renewable energy applications. Energies 9 https://doi.org/10.3390/en9080607. <http://www.mdpi.com/1996-1073/9/8/607>.
Peruchena, C.F., Larrañeta, M., Blanco, M., Bernardos, A., 2018. High frequency generation of coupled GHI and DNI based on clustered dynamic paths. Sol. Energy 159, 453–457. https://doi.org/10.1016/j.solener.2017.11.024. <http://www.sciencedirect.com/science/article/pii/S0038092X17310071>.
Peruchena, C.M.F., Blanco, M., Gastón, M., Bernardos, A., 2015. Increasing the temporal resolution of direct normal solar irradiance series in different climatic zones. Sol. Energy 115, 255–263. https://doi.org/10.1016/j.solener.2015.02.017. <http://www.sciencedirect.com/science/article/pii/S0038092X15000870>.
Polo, J., Zarzalejo, L., Marchante, R., Navarro, A., 2011. A simple approach to the synthetic generation of solar irradiance time series with high temporal resolution. Sol. Energy 85, 1164–1170. https://doi.org/10.1016/j.solener.2011.03.011. <http://www.sciencedirect.com/science/article/pii/S0038092X11000946>.
Rasmussen, C.E., 1999. The infinite Gaussian mixture model. In: Proceedings of the 12th International Conference on Neural Information Processing Systems. MIT Press, Cambridge, MA, USA, pp. 554–560.
Schallenberg-Rodríguez, J., Montesdeoca, N.G., 2018. Spatial planning to estimate the offshore wind energy potential in coastal regions and islands. practical case: The canary islands. Energy 143, 91–103. https://doi.org/10.1016/j.energy.2017.10.084. <http://www.sciencedirect.com/science/article/pii/S0360544217317899>.
Schwarz, M., Folini, D., Hakuba, M.Z., Wild, M., 2018. From point to area: Worldwide assessment of the representativeness of monthly surface solar radiation records. J. Geophys. Res.: Atmosph. 123, 13,857–13,874. https://doi.org/10.1029/2018JD029169.
Sethuraman, J., 1994. A constructive definition of Dirichilet priors. Statistica Sinica 4, 639–650.<http://www.jstor.org/stable/24305538>.
Shi, X., Acord, B., Wang, P., 2018. Incorporating ground-measured pollution observations to improve temporally downscaled solar irradiance simulations. Sol. Energy 171, 293–301. https://doi.org/10.1016/j.solener.2018.06.076. <http://www.sciencedirect.com/science/article/pii/S0038092X18306273>.
Smith, C.J., Bright, J.M., Crook, R., 2017. Cloud cover effect of clear-sky index distributions and differences between human and automatic cloud observations. Sol. Energy 144, 10–21. https://doi.org/10.1016/j.solener.2016.12.055. <http://www.sciencedirect.com/science/article/pii/S0038092X16306624>.
Soubdhan, T., Ndong, J., Ould-Baba, H., Do, M.T., 2016. A robust forecasting framework based on the Kalman filtering approach with a twofold parameter tuning procedure: Application to solar and photovoltaic prediction. Sol. Energy 131, 246–259. https://doi.org/10.1016/j.solener.2016.02.036. <http://www.sciencedirect.com/science/article/pii/S0038092X16001444>.
Sun, Y., Wong, A.K.C., Kamel, M.S., 2009. Classification of imbalanced data: A review. Int. J. Pattern Recognit Artif Intell. 23, 687–719. https://doi.org/10.1142/S0218001409007326.
Tomson, T., Tamm, G., 2006. Short-term variability of solar radiation. Sol. Energy 80, 600–606. https://doi.org/10.1016/j.solener.2005.03.009. <http://www.

sciencedirect.com/science/article/pii/S0038092X05001398>.
Wang, F., Zhen, Z., Mi, Z., Sun, H., Su, S., Yang, G., 2015. Solar irradiance feature extraction and support vector machines based weather status pattern recognition model for short-term photovoltaic power forecasting. Energy Build. 86, 427–438. https://doi.org/10.1016/j.enbuild.2014.10.002.
Yang, D., Jirutitijaroen, P., Walsh, W.M., 2012. Hourly solar irradiance time series forecasting using cloud cover index. Sol. Energy 86, 3531–3543. https://doi.org/10.1016/j.solener.2012.07.029. solar Resources. <http://www.sciencedirect.com/science/article/pii/S0038092X12003039>.
Zagouras, A., Pedro, H.T., Coimbra, C.F., 2014. Clustering the solar resource for grid management in island mode. Sol. Energy 110, 507–518. https://doi.org/10.1016/j.solener.2014.10.002. <http://www.sciencedirect.com/science/article/pii/S0038092X14004836>.
Zhang, W., Kleiber, W., Florita, A.R., Hodge, B.M., Mather, B., 2018. A stochastic downscaling approach for generating high-frequency solar irradiance scenarios. Sol. Energy 176, 370–379. https://doi.org/10.1016/j.solener.2018.10.019. <http://www.sciencedirect.com/science/article/pii/S0038092X1830999X>.
Zhong, X., Kleissl, J., 2015. Clear sky irradiances using REST2 and MODIS. Sol. Energy 116, 144–164. https://doi.org/10.1016/j.solener.2015.03.046. <http://www.sciencedirect.com/science/article/pii/S0038092X15001735>.