Journal of Renewable and Sustainable Energy

A generative hidden Markov model of the clear-sky index

Cite as: J. Renewable Sustainable Energy **11**, 043703 (2019); https://doi.org/10.1063/1.5110785
Submitted: 21 May 2019 . Accepted: 13 July 2019 . Published Online: 31 July 2019

Mahmoud Shepero, Joakim Munkhammar, and Joakim Widén

Journal of Renewable and Sustainable Energy ARTICLE scitation.org/journal/rse

A generative hidden Markov model of the clear-sky index

Cite as: J. Renewable Sustainable Energy **11**, 043703 (2019); doi: 10.1063/1.5110785
Submitted: 21 May 2019 · Accepted: 13 July 2019 ·
Published Online: 31 July 2019

Mahmoud Shepero,[a] Joakim Munkhammar,[b] and Joakim Widén[c]

AFFILIATIONS

Department of Engineering Sciences, Uppsala University, P.O. Box 534, SE-751 21 Uppsala, Sweden

[a]Electronic mail: mahmoud.shepero@angstrom.uu.se
[b]Electronic mail: joakim.munkhammar@angstrom.uu.se
[c]Electronic mail: joakim.widen@angstrom.uu.se

ABSTRACT

Clear-sky index (CSI) generative models are of paramount importance in, e.g., studying the integration of solar power in the electricity grid. Several models have recently been proposed with methodologies that are related to hidden Markov models (HMMs). In this paper, we formally employ HMMs, with Gaussian distributions, to generate CSI time-series. The authors propose two different methodologies. The first is a completely data-driven approach, where an HMM with Gaussian observation distributions is proposed. In the second, the means of these Gaussian observation distributions were predefined based on the fraction of time of bright sunshine from the site. Finally, the authors also propose a novel method to improve the autocorrelation function (ACF) of HMMs in general. The two methods were tested on two datasets representing two different climate regions. The performance of the two methodologies varied between the two datasets and among the compared performance metrics. Moreover, both the proposed methods underperformed in reproducing the ACF as compared to state-of-the-art models. However, the method proposed to improve the ACF was able to reduce the mean absolute error (MAE) of the ACF by up to 19%. In summary, the proposed models were able to achieve a Kolmogorov-Smirnov test score as low as 0.042 and MAE of the ACF as low as 0.012. These results are comparable with the state-of-the-art models. Moreover, the proposed models were fast to train. HMMs are shown to be viable CSI generative models. The code for the model and the simulations performed in this paper can be found in the GitHub repository: *HMM-CSI-generativeModel*.

I. INTRODUCTION

A. Literature overview

The variability of solar irradiance on the Earth's surface affects many solar engineering applications such as photovoltaic (PV) power generation.[1–3] By quantifying the solar irradiance variability, the design and operation of power systems with large amounts of grid-connected distributed PV power generation can be improved.[2,4] However, quantifying and reproducing the complexity of solar irradiance variability over time is a challenge for statistical and machine learning models.[5] The value of high resolution irradiance data might increase in the future.[4,6] Consequently, several recent models have been proposed to generate high resolution irradiance data using low resolution data.[7–10]

The solar irradiance is often normalized using the clear-sky irradiance, producing the clear-sky index (CSI). The CSI has interesting features of statistical complexity, particularly on minute to instantaneous scale.[11,12] High resolutions are favorable in order to take short bursts of overirradiance into account,[13] but resolutions above 1 s are arguably less informative and only increase the complexity of data management.[14]

There exist irradiation estimates from ground measurements or from satellite data, see, e.g., Ref. 15, and from software, such as Meteonorm.[16] Despite this, more high resolution data are needed for many locations,[11] and when data are insufficient, realistic synthetic data generated by generative models are useful.[5] Also, when modeling CSI variability, it is of importance to quantify the model output accuracy so that it generates realistic data.[5,17–20] In terms of modeling, the distribution of minute to instantaneous CSI can be modeled by two,[20,21] three,[22,23] n-single peak distributions,[5] or n-multipeak distributions.[24]

In order to ensure that the model generates CSI samples that resemble that of the training CSI data, models are evaluated based on first distribution goodness-of-fit and second autocorrelation function (ACF) likeness. Model evaluation and selection, in terms of output probability distribution goodness-of-fit, are typically measured using, e.g., Kolmogorov-Smirnov (K-S) statistics.[12,17,25–27] ACF likeness compares the temporal variability between models and data, and the ACF

measured over a set of lags has commonly been used.[17,19,20] A metric for model selection is, for example, a mean absolute error (MAE) over lags in the ACF.[5]

Models aiming to quantify the CSI and generate realistic synthetic CSI data, with proper ACF likeness, include Gaussian-Markov,[19] auto-regressive Gaussian,[28] neural networks,[29] copula modeling,[17,26] fractal cloud modeling,[30] Markov chains,[5,12,18,31–34] and Dirichlet process Gaussian mixture (DPGM).[24] These CSI generators are practical since they utilize some existing dataset of lower resolution or averaged CSI data to estimate higher resolution data (temporal or spatial), see, e.g., Refs. 12, 25, 31, and 34–36, or smaller amounts of CSI data to generate unlimited amounts of data.[5,17,18,26]

In particular, Markov chains, or generally Markov models, have been useful as CSI generators, where Ref. 34 generated minute resolutions, while in Ref. 32, the daily resolution was modeled. In Ref. 18, a two-state Markov-chain mixture probability distribution was used, similar to the one in the study of Morf,[31] where a general model for generating clear and cloudy periods was constructed. As a generalization of the two-state models, an n-state piecewise uniform Markov-chain mixture distribution model was developed in Refs. 5 and 10, which had low model complexity, yet high accuracy. Generally, these models have presented high accuracy in both distribution goodness-of-fit and ACF likeness.

A novel approach was recently proposed by Frimane *et al.*[24] where a DPGM model was employed to generate CSI samples. Unlike previous models, the DPGM model is a nonparametric model that selects the number of CSI clusters based on the training data. The daily distribution of CSI is considered to be represented by an infinite mixture of Gaussian distributions. A Dirichlet process was used to cluster the daily distribution of CSI into a number of clusters, with some probability of creating a new cluster for the observed daily CSI. Thus, the number of clusters was inferred from the data. A Markov chain was then learned from the data, which was used to generate the CSI time-series.

One conclusion from the literature is that mixture models based on stochastic processes (Markov-chain or otherwise) reproduce the CSI variability comparatively well and that models, which can utilize meteorological observables and produce realistic output with high accuracy, are particularly useful.

An improved understanding regarding three-state modeling connected to meteorological variables,[22] in particular, in combination with hidden Markov models (HMMs), has been proposed but not yet investigated in the literature.[18] Also, it can be concluded that HMMs with Gaussian distributions have not been investigated in the literature either. In this paper, both a three-state Gaussian HMM model, with connections to the measured fraction of time of bright sunshine, and a general machine learning-based HMM for n states are developed and investigated.

The developed models are beneficial to generate CSI time-series in locations of interest. The generated time-series can then be converted to the global irradiance, which can be used in applications such as PV potential studies.

B. Contribution

This study aims to develop an HMM for the CSI based on n Gaussian distributions. The model extends previous Markov-chain mixture distribution models[5,18] and previous literature through

1. Developing an HMM generative model for the CSI. Previous contributions did not explicitly employ HMMs.
2. Connecting the HMM model to previous meteorological CSI models.
3. Proposing a methodology to improve the ACF of the fitted HMMs; to the best of the author's knowledge, this method has not been proposed before in the literature.

The model is trained and tested on solar irradiance datasets for Hawaii and Norrköping, which were also used in Refs. 5, 17, and 18.

This paper is organized as follows. The proposed HMMs are described in Sec. II. In Sec. III, the results obtained from implementing the model are presented. A discussion is provided in Sec. IV. Finally, conclusions are drawn in Sec. V.

II. METHODS

A general introduction to HMMs is provided in Sec. II A. In Sec. II B, the HMM is applied to CSI modeling. This section connects HMMs to CSI observations in a certain location. Section II C presents the two proposed methods to employ HMMs in CSI time-series generation. Performance metrics, used in comparing the various models, are described in Sec. II D. Finally, data used in learning and comparing the models are introduced in Sec. II E.

A. Hidden Markov model

Here, a brief introduction to HMMs is provided. For more detailed information regarding HMMs, Refs. 37 and 38 can be consulted.

A discrete-time HMM is a state space model that is characterized by a set of observations $\{X_t\}_{t=1}^{T}$. These observations were observed at discrete time-steps $t \in \{1, 2, \ldots, T\}$. Furthermore, the observations are dependent on a set of hidden states $\{S_t\}_{t=1}^{T}$ such that the observations are conditionally independent given the states, see Fig. 1.

The transition probabilities of the hidden states follow a Markov chain. A Markov chain is a memoryless process such that the hidden state S_t is only dependent on the previous hidden state S_{t-1} and independent of the previous state trajectory, i.e.,

$$P(S_t = j|S_1, \ldots, S_{t-1} = i) = P(S_t = j|S_{t-1} = i) = p_{ij}, \quad (1)$$

where p_{ij} is the probability of transitioning from state i to state j. A transition matrix A between the states can be formed

$$A = \begin{bmatrix} p_{11} & \cdots & p_{1n} \\ \vdots & \ddots & \vdots \\ p_{n1} & \cdots & p_{nn} \end{bmatrix}, \quad (2)$$

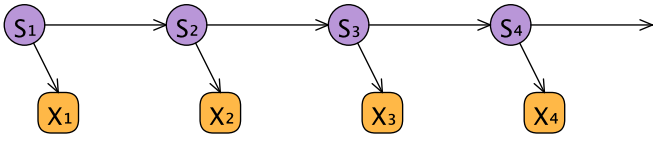

FIG. 1. A graphical representation of the HMM for four time-steps. The arrows represent conditional dependencies. The hidden states $\{S_1, S_2, \ldots\}$ are represented by circles and the observations $\{X_1, X_2, \ldots\}$ by squares.

where n is the number of hidden states, i.e., the cardinality of $\{S_t\}_{t=1}^{T}$. The transition matrix A has to satisfy

$$\sum_{j=1}^{n} p_{ij} = 1, \tag{3}$$

for every initial state i.

The observations X_t are only dependent on the hidden state at time t,

$$\begin{aligned} &P(X_t = x_t | S_1, \ldots, S_t = i, X_1, \ldots, X_{t-1}) \\ &\quad = P(X_t = x_t | S_t = i) = b_i(x_t), \end{aligned} \tag{4}$$

where $b_i(x_t)$ is the observation distribution of state i, e.g., a Gaussian distribution.

The HMM is characterized by the model parameters θ. The model parameters θ consist of the initial distribution of the hidden states, the transition matrix A, and the parameters of the observation distributions $b_i(x_t)$. The parameters $b_i(x_t)$ can be, for example, the mean μ_i and variance σ_i^2 for univariate Gaussian observations.

Assuming that the hidden process is aperiodic and irreducible and assuming that the process is stationary, the initial distribution can be replaced by the stationary distribution $\boldsymbol{\pi}$ of the Markov chain, see, e.g., Ref. 39, p. 66. In other words, if the process has been going on for a long time, e.g., the CSI time-series started long ago, the initial distribution loses importance and the stationary distribution becomes more important. The stationary distribution $\boldsymbol{\pi}$ can be estimated by solving

$$\boldsymbol{\pi} A = \boldsymbol{\pi}. \tag{5}$$

In HMMs, the model parameters θ are learned from sequences of observations using the Baum-Welch algorithm. The Baum-Welch algorithm is of computational complexity $\mathcal{O}(n^2 T)$, see Ref. 37, where n is the number of hidden states and T is the length of the training time-series.

B. Clear-sky index as HMM observations

This section draws connections between HMMs and CSI observations in a certain location. In other words, this section shows that the CSI time-series can be viewed as the observation of an HMM.

Assume that the hidden states S_t of the HMM represent the status of the cloud cover in the sky at every time step t in a certain location. Following Hollands and Suehrcke,[22] the cloud cover can be categorized into three categories, $S \in \{1, 2, 3\}$. These categories can be conceptualized as states of the atmosphere.[22] For clarity, we propose, here, to call these hidden states obscured (1), unobscured in clear sky (2), and unobscured in partially cloudy sky (3). Consequently, the transition matrix A is a (3×3)-matrix.

As shown in Fig. 2, each hidden state is coupled with an observation distribution, defined by a probability density function (PDF). In such a case, the CSI measurements in a certain location can be conceptualized as observations from the hidden—or unobserved—states in the HMM model. These hidden states represent the cloud cover in the sky.

Pursuing the same derivation as in Ref. 18, few model parameters can be extracted from meteorological measurements. The probability of having an obscured sun for δ_1 consecutive time steps—then switching to a different state—is (Ref. 40, p. 308)

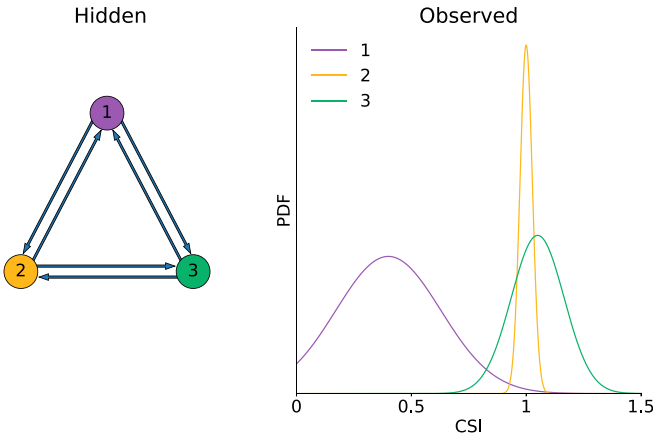

FIG. 2. A diagram connecting the hidden states of a three state HMM—representing the cloud cover states—and the CSI. The hidden states are represented by the circles, and the between-state transitions are represented by arrows. The remain-in-state transitions are not presented. The observed distribution of every state is plotted on the right. State 1 represents the obscured state, state 2 represents the unobscured in the clear-sky state, and state 3 represents the unobscured in partially cloudy sky.

$$P(\delta_1) = (1 - p_{11}) p_{11}^{\delta_1 - 1}. \tag{6}$$

The expected value of obscured durations $\mathbb{E}[\delta_1]$ can be estimated as[18,40]

$$\mathbb{E}[\delta_1] = \frac{1}{1 - p_{11}}. \tag{7}$$

Applying Eq. (7) for all the three hidden states results in

$$\begin{aligned} p_{11} &= \frac{\mathbb{E}[\delta_1] - 1}{\mathbb{E}[\delta_1]}, \\ p_{22} &= \frac{\mathbb{E}[\delta_2] - 1}{\mathbb{E}[\delta_2]}, \\ p_{33} &= \frac{\mathbb{E}[\delta_3] - 1}{\mathbb{E}[\delta_3]}, \end{aligned} \tag{8}$$

where $\mathbb{E}[\delta_2]$ and $\mathbb{E}[\delta_3]$ are the mean durations of unobscured in clear-sky and unobscured in partially cloudy sky, respectively. These can be directly observed or calculated from other measured values in a certain location, see, e.g., Refs. 18, 22, 31, 41, and 42.

Hollands and Suehrcke[22] proposed Gaussian distributions for each hidden state. The means of the Gaussian distributions were proposed to be $\mu_2 = 1$, $\mu_3 = 1.04$, and

$$\mu_1 = 0.1205 + 0.3341\mathcal{K}, \tag{9}$$

where $\mathcal{K}$ is the ratio between the mean CSI in the location and the CSI at the peak of the distribution. Here and as proposed in Refs. 22 and 41, we calculate $\mathcal{K} = \sqrt{\tau}$, where τ is the fraction of time of bright sunshine. Here, τ is defined as the fraction of time where the CSI is larger than 0.95, see Ref. 18.

The HMM has further analytical properties in regard to the temporal variability, e.g., the ACF $\rho(k)$ for the k-th time-lag of the observed CSI time-series X_t is defined as[43]

$$\rho(k) = \frac{\mathbb{E}[(X_t - \mu)(X_{t+k} - \mu)]}{\sigma^2}, \tag{10}$$

where μ and σ^2 represent the mean and variance of the time-series. In HMMs, Eq. (10) can be shown to follow (Ref. 40, p. 310)

$$\rho(k) = \frac{\sum_i \pi_i \mu_i \sum_j \mu_j (A^k)_{ij} - \mu^2}{\sigma^2}, = \frac{\boldsymbol{\pi}\mathrm{diag}(\boldsymbol{\mu}) A^k \boldsymbol{\mu}' - \mu^2}{\sigma^2}, \tag{11}$$

where $(A^k)_{ij}$ is the value in the ith row and the jth column of the A^k matrix, $\boldsymbol{\mu}$ is a vector containing the means of the observation distributions of the hidden states; and μ and σ^2 are the mean and variance of the mixture distribution of the observations, respectively. In the case of Gaussian mixture distributions (Ref. 40, p. 11),

$$\mu = \sum_{i=1}^{n} \pi_i \mu_i \tag{12}$$

and

$$\sigma^2 = \sum_{i=1}^{n} \pi_i (\mu_i^2 + \sigma_i^2) - \mu^2. \tag{13}$$

Equation (11) is applicable only for $k > 0$ (Ref. 39, p. 70). For $k = 0$, $\rho(0) = 1$ by definition.[44] One limitation of the HMMs is that they can only produce exponentially decaying ACFs[45] and observe that $\rho(k) \propto A^k$ for all $k > 0$ in Eq. (11).

An HMM with n hidden states and univariate Gaussian observations can be defined by $(n^2 + 2n)$ parameters; n^2 parameters in A, n parameters in $\boldsymbol{\mu}$, and n parameters in $\boldsymbol{\sigma}^2$ which is the vector containing the variances of the n observation distributions.

Fitting a Gaussian mixture model to the CSI distribution provides an estimate of $\boldsymbol{\mu}$, $\boldsymbol{\sigma}^2$, and $\boldsymbol{\pi}$, see, e.g., Refs. 22 and 23. The stationary distribution of the Markov process $\boldsymbol{\pi}$ controls the contribution of each Gaussian observation distribution to the CSI mixture distribution.

In order to estimate the transition matrix A, Eqs. (5), (8), (3), and (11) can, theoretically, be employed. However, Eq. (11) is nonlinear, because of the term A^k. As a result, the complexity grows as the number of hidden states grows. For example, for an HMM with $n = 2$, Munkhammar and Widén[18] showed that solving Eqs. (3) and (8) was sufficient to estimate A.

For an HMM with $n = 3$, Eqs. (3), (8), and (5) provide 8 independent equations as the 9th equation can be derived using the remaining 8 equations. Equation (11) can provide the 9th independent equation, however, only if $k > 1$ since the equation for $\rho(1)$ can be derived from the previously mentioned 8 equations. Consequently, a nonlinear equation solving method is needed if this approach is to be pursued.

C. Model implementation

We assume that the hidden states of an HMM represent some hidden, or unobserved, cloud cover states. These cloud cover states in turn control the CSI observations. Using the CSI observations, an HMM with various numbers of states can be fitted using the Baum-Welch algorithm. These models will be called the Gaussian hidden Markov models (GHMMs) in further discussions since the observation distributions for each hidden state are assumed to be Gaussian. The Gaussian distribution was chosen since it was shown to adequately represent the CSI time-series in Refs. 22 and 24. Selecting the optimal number of hidden states n will be based on comparing the fitted models using some performance metrics.

The GHMM with $n = 2$ shares some similarities with the models proposed in Refs. 18, 31, and 46. Here, unlike Ref. 18, the observation distributions are Gaussian, and they are not truncated at a middle point. In comparison, Munkhammar and Widén[18] used Gaussian, log-normal and polynomial distributions in their two-state model.

It is important to note that in all HMMs trained here, the Baum-Welch algorithm was employed to estimate the model parameters θ. The Baum-Welch algorithm updates the model parameters θ in each iteration such that θ increases the log-likelihood of the model. The final model is thus the model that locally maximizes the log-likelihood. As a result, for the GHMM with $n = 3$, the estimated means of the observation distributions of the hidden states might not represent the cloud cover assumptions proposed by Hollands and Suehrcke.[22]

An alternative approach to the GHMM with $n = 3$ is to fix the means of the observation distributions and thereby connect it to the cloud cover assumptions of Hollands and Suehrcke.[22] In this alternative model, the Baum-Welch algorithm was employed to estimate all the model parameters θ except the means of the observation distributions $\boldsymbol{\mu}$. The first mean μ_1 was predefined according to Eq. (9), and μ_2 and μ_3 were set to 1.0 and 1.04, respectively.[22] This alternative modeling approach will be called the fixed-means Gaussian hidden Markov model (FMGHMM) in further discussions.

In similarity with Ref. 5, this paper uses HMMs with $n > 3$, but unlike Ref. 5, Gaussian, instead of uniform, observation distributions are employed. A nonparametric model for the number of hidden states was recently developed in Ref. 24. In this model, the number of hidden states was estimated from the training data. Thus, it can be understood as a GHMM with an infinite number of hidden states.[47] Nevertheless, the authors used mixed Gaussian observation distributions.

To summarize this section, two modeling strategies were proposed: the GHMMs and the FMGHMMs. Using the GHMMs, the number of hidden states and the shapes of the observation distributions were learned completely from the data. Consequently, the hidden states of the GHMMs are not guaranteed to reflect the assumptions of Hollands and Suehrcke.[22] On the other hand, the FMGHMM used the assumptions in Ref. 22 to fix the means of the observation distributions. This makes the FMGHMM more connected to the theoretical properties of the CSI. The FMGHMM is, thus, only meaningful if the number of hidden states $n = 3$.

In this paper, the Python HMM package `hmmlearn v0.2.1`[48] was used to fit the proposed models.

D. Model evaluation metrics

Theis *et al.*[49] compared the various metrics used in evaluating generative models, e.g., HMMs. The authors concluded that generative models need to be evaluated with respect to their applications.

Consequently, we propose the following performance metrics to compare the proposed models:

1. Log-likelihood $\mathcal{L}$ of the training and test datasets.
2. K-S test score K between both the training and test datasets and samples from the model.

3. MAE of the ACF ϵ between the test data and samples from the model.
4. Training time.

The log-likelihood, $\mathcal{L} = \log(P(X_1, \ldots, X_T|\theta)$, represents the natural logarithm of the probability of observing the CSI time-series given the model parameters θ. This metric was measured for both the training and the test datasets.

Following the conclusions of Theis *et al.*[49] and to evaluate the models with regard to their applications, two metrics were added to measure first the goodness-of-fit of the model distribution and the distribution of the CSI time-series and second the ACF likeness between the generated samples and the ACF of the CSI time-series. As stated before, these two metrics were commonly used in the literature. The K-S test score was employed before in Refs. 5, 12, 17, 18, and 25–27. The MAE of the ACF was employed before in Refs. 5 and 18. Consequently, the chosen metrics enable comparing the proposed methods with several state-of-the-art methods. Future studies can, nonetheless, test the proposed models in regard to different error metrics.

To evaluate the goodness-of-fit of the distribution, the K-S test was used. The K-S test score K evaluates the similarity between two distributions

$$K = \max|F_1(x) - F_2(x)|, \tag{14}$$

where $F_1(x)$ and $F_2(x)$ are the empirical cumulative distribution functions of the two compared distributions.[50]

To evaluate the ACF likeness, the MAE of the ACF ϵ was evaluated, which is estimated as

$$\epsilon = 1/60 \sum_{k=1}^{60} |\hat{\rho}(k) - \rho(k)|, \tag{15}$$

where $\hat{\rho}(k)$ is the ACF of the test data and $\rho(k)$ is the ACF of the samples from the model. The ACF error was calculated for one hour, i.e., 60 time-lags. This makes our ϵ results comparable with those of Refs. 5 and 18.

The reason for only recording the MAE of the ACF ϵ on the test data is that by visual inspection, as will be shown later in Figs. 5, 8, and 11, there was no difference between the ACF of the training and test data for both locations.

Finally, the training time was measured in seconds.

E. Data

The data used to train and evaluate the models were based on radiometer measurements of global horizontal irradiance (GHI) for one year obtained from the Swedish Meteorological and Hydrological Institute (SMHI) for Norrköping, Sweden (59°35′31″ N 17°11′8″ E),[51] and from the National Renewable Energy Laboratory (NREL) radiometer array in Oahu, Hawaii, USA (21°31′ N, 158°09′ W).[52] The data were recorded during the years 2008 and 2010 for Norrköping and Hawaii, respectively. The Kipp and Zonen CM 21[53] and LI-COR LI-200[54] pyranometers were used in Norrköping and Hawaii, respectively.

Data from both the locations represent instantaneous irradiance with one minute resolution, averaged from raw data recorded with higher resolutions. From these datasets, 120 data points (minutes) per day (centered around noon each day) for both the locations were used in order to avoid low solar angles. Thus, the selected datasets totaled 43 800 (120×365) data points from each location. These particular datasets were previously selected for comparative reasons since they were also used in Refs. 17 and 26 to generate an autocorrelation model of the CSI using copulas and in Refs. 5 and 18 for the Markov-chain mixture distribution models. Testing the performance of the proposed methods on data from different locations, and representing various solar altitude angles, climate conditions, air masses, etc., is encouraged for future works.

The CSI normalization for the Norrköping data was made with the Ineichen-Perez clear-sky irradiance model[55] and for the Hawaii data the McClear clear-sky irradiance model[56] was used. The use of these clear-sky irradiance models for obtaining the CSI for both the datasets was described, tested, and found to be optimal in Refs. 17 and 26. Thus, these particular clear-sky irradiance models were chosen in order to enable comparisons with previous papers that used the same datasets. Moreover, the two clear-sky irradiance models had similar performances on the datasets, which further motivated using both models to enable comparison of results with previous studies. For a detailed review on the recommended clear-sky irradiance models for each climate zone, Ref. 57 can be consulted.

The CSI data from both locations were divided into training data and test data. The division was made on an every-other-day basis in order to minimize the seasonality effect on training of the models. In other words, the training data were the data recorded on the days 1, 3,…,365; and the test data were the data from the days 2, 4,…,364.

For each scenario, the model was used to generate a synthetic time-series with a length of 1 million, i.e., $T = 10^6$. This sample is large enough to ensure stable estimates of the performance metrics of the models.

III. RESULTS

This section provides the results of fitting the proposed HMMs to the CSI data. Section III A provides the results of the GHMMs. In Sec. III B, the results of the FMGHMM are presented. Section III C improves the ACF of the fitted HMMs, with the FMGHMM taken as an example. In Sec. III D, our results are compared with the results of some previously proposed models.

A. Gaussian hidden Markov models

In this section, the performance results of the GHMMs are presented. The Baum-Welch algorithm was employed to learn all the model parameters θ including the means of the observation distributions $\boldsymbol{\mu}$—unlike the FMGHMM results to be presented in Sec. III B.

As shown in Figs. 3(a) and 3(b), the log-likelihood $\mathcal{L}$ of the training and the test datasets increased as the number of hidden states increased. The rate of increase in the log-likelihood, however, decreased as the number of hidden states increased. This indicated that the improvement increments in model fitting to the data were high for the models with a lower number of hidden states.

The K-S test scores K from the models were lower than 0.1 for both locations and for both the training and test datasets, see Figs. 3(c) and 3(d). The K-S test score was 0.050 for both the training and test datasets for $n = 3$ in the case of Hawaii. In the case of Norrköping, the scores were 0.047 and 0.048 for the training and test datasets, respectively. The minimum K-S test score on the test dataset was 0.008 at $n = 9$, for the Hawaii case, and 0.012 at $n = 9$, for the Norrköping case.

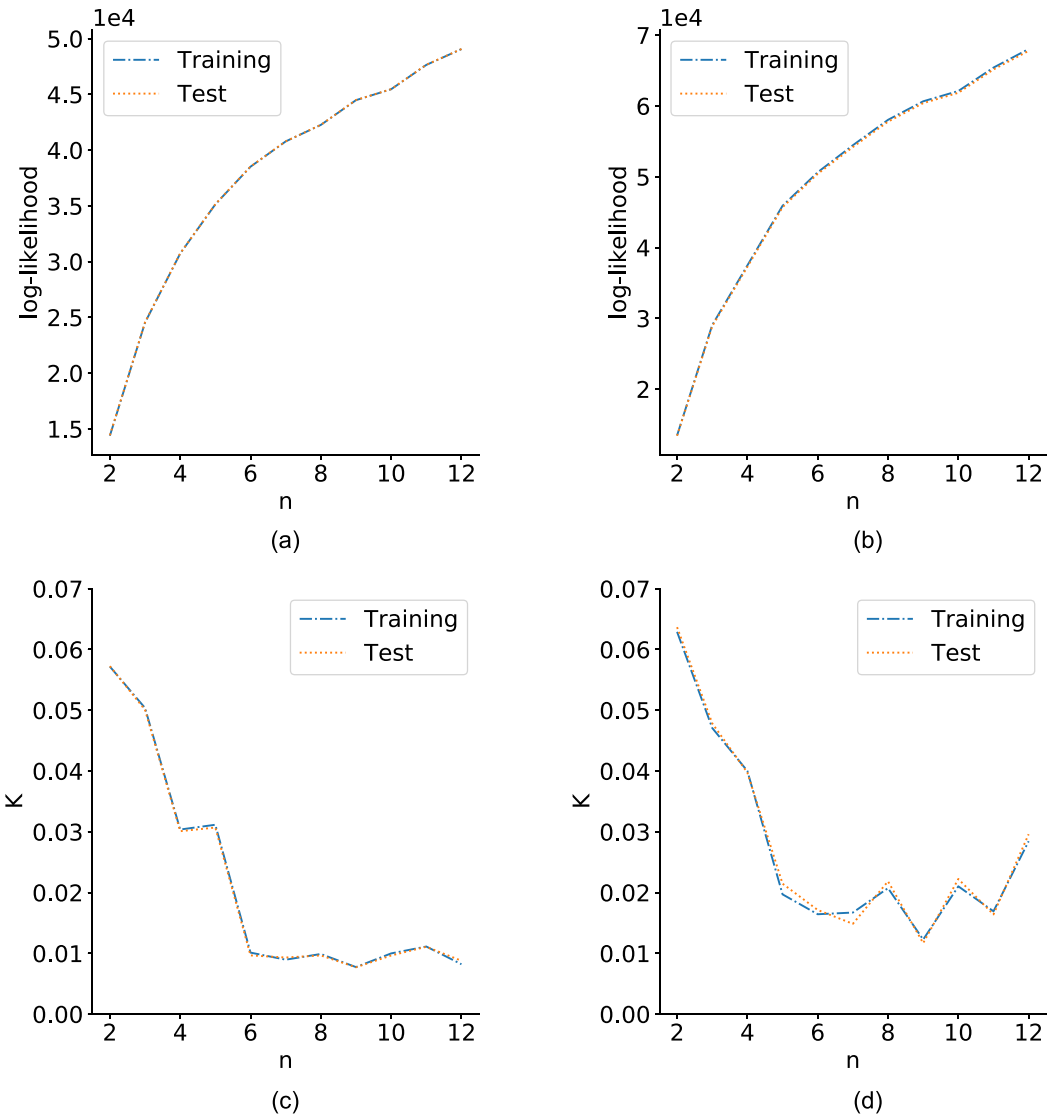

FIG. 3. The log-likelihood $\mathcal{L}$, in (a) and (b), and the K-S test score K, in (c) and (d), results for the GHMMs with different numbers of hidden states n. (a) and (c) present the results of Hawaii. (b) and (d) present the results of Norrköping.

Figures 4(a) and 4(b) compare the PDFs of the samples from the fitted GHMMs to the test data. The Python class `scipy.stats.gaussian_kde` was used, with the default parameters, to smoothen the histograms and perform a kernel density estimation (KDE). Still, the exact histograms for both the locations are presented in Figs. 4(c) and 4(d).

Figures 4(a) and 4(c) show that for Hawaii, the model with two hidden states, $n = 2$, under-represented the CSI peak at CSI ≈ 1. In addition, the first mode—peak—of the bimodal CSI distribution was situated at a higher CSI compared to the test data, 0.45 and 0.333, respectively. For $n = 3$, the model results seem to match the cloud cover assumptions of Hollands and Suehrcke.[22] Moreover, the PDF of the model samples decently represented the second mode of the test data. Still, the model estimated the first mode of the distribution at a higher CSI, 0.433 compared to 0.333 for the test data. The exact closeness scores between the distributions are presented in Figs. 3(c) and 3(d).

For the Norrköping data, the CSI distribution was more skewed toward the obscured cloud cover states, see Figs. 4(b) and 4(d). The first mode value of the PDF was at CSI = 0.19 for the test data. The GHMM with $n = 2$ under-represented the mode at CSI ≈ 1. In addition, the first mode of the PDF was estimated at CSI = 0.33. For $n = 3$, the model allocated two distributions for the first half of the PDF. This improved the fit to the obscured part of the PDF but still under-represented the mode at CSI ≈ 1. The model with $n = 5$ adequately represented the test data. This model allocated two hidden states each to the two unobscured-sun states and three hidden states to represent the obscured-sun state.

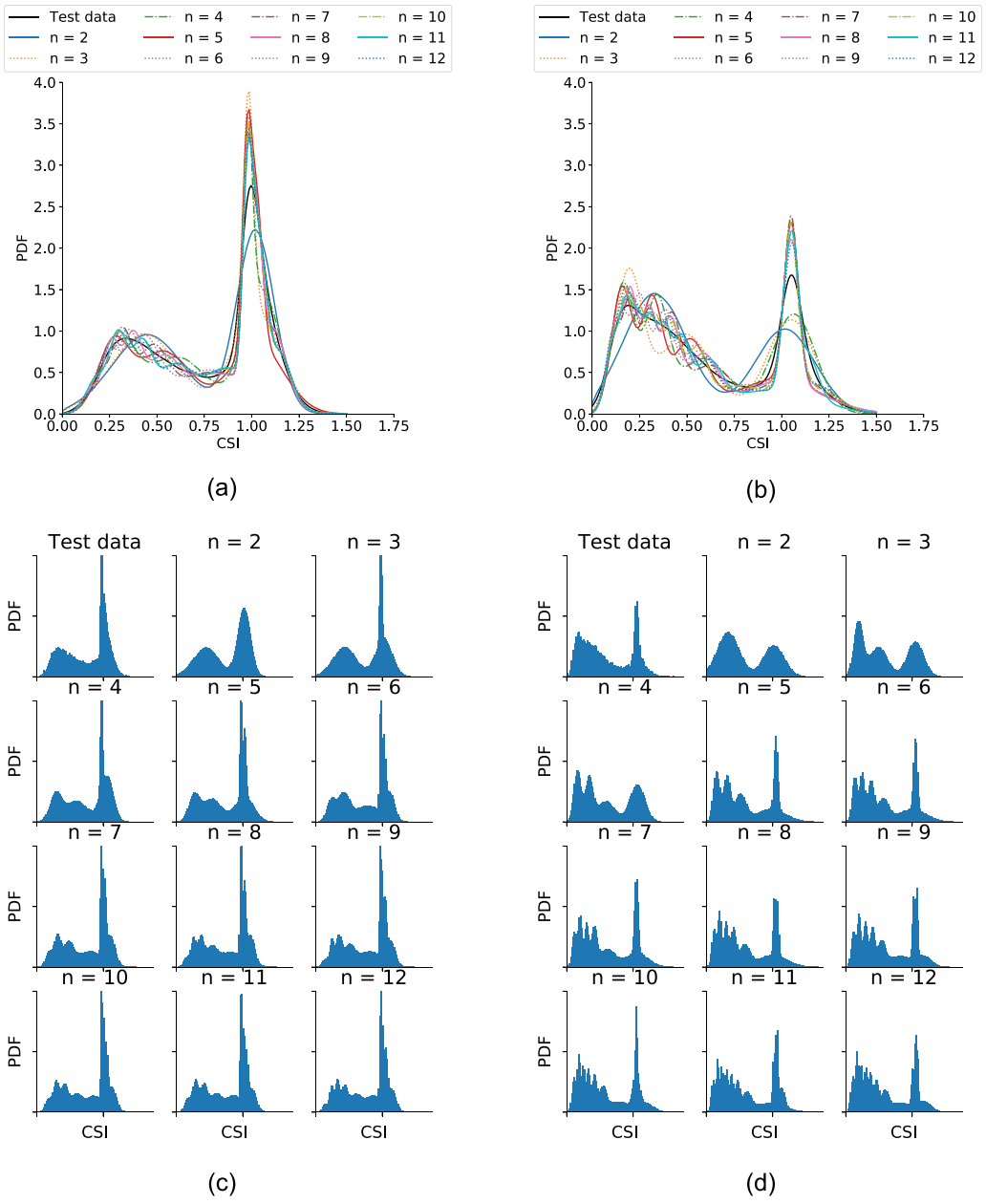

FIG. 4. Kernel density estimation (KDE), in (a) and (b), and Histograms, in (c) and (d), of the CSI of the test data and samples from the GHMMs with different numbers of hidden states *n*. (a) and (c) present the results of Hawaii. (b) and (d) present the results of Norrköping.

In order to facilitate further comparisons with the FMGHMM, the results of the GHMM with $n = 3$ are summarized in Tables I and II. Table I presents the fitted observation distributions of the GHMM with three hidden states for both locations. For the Hawaii dataset, the distributions resembled the three clusters from Ref. 22, presented here in Sec. II B. In Norrköping, however, the HMM combined the two unobscured-sun states—from Sec. II B—into one state, state 3 in Table I. The obscured-sun state was represented by the two remaining distributions. Table II presents the performance metrics of the GHMM with $n = 3$.

In regard to the ACF, the GHMM did not accurately represent the ACF of neither the training nor the test data—both had similar ACFs—see Fig. 5. For Hawaii, the ACF from model samples did not improve as the number of hidden states increased. For Norrköping, the ACF of samples from the models with $n > 3$ closely resembled that of the training and test data.

TABLE I. The observation distributions for the GHMM model with $n = 3$ for both locations. All distributions are Gaussians.

	Hawaii		Norrköping	
State	μ_i	σ_i^2	μ_i	σ_i^2
1	0.433	0.0312	0.190	0.0048
2	0.982	0.00032	0.490	0.019
3	1.025	0.016	1.05	0.0154

TABLE II. The performance metrics for the GHMM for both locations and datasets for $n = 3$. Note that the MAE of the ACF ϵ was only estimated on the test dataset. On the other hand, the log-likelihood $\mathcal{L}$ and the K-S test score K were estimated for both datasets.

	Training		Test		
Location	$\mathcal{L}$	K	$\mathcal{L}$	K	ϵ
Hawaii	24534.25	0.050	24557.19	0.050	0.24
Norrköping	29058.64	0.047	28893.89	0.048	0.16

The MAE of the ACF ϵ is presented in Fig. 6(a). For Hawaii, ϵ stagnated at approximately 0.2 or 20%. This was perhaps expected given that Rydén *et al.*[45] also noticed that their HMM failed to properly represent the ACF of the data. The authors mentioned that models fitted with maximizing the likelihood cannot accurately capture the profile of the ACF. A method is proposed and tested in Sec. III C to improve the ACF of HMMs.

The training time of different models is presented in Fig. 6(b). As expected from the Baum-Welch algorithm, the training time increased with the square of the number of states.

B. Fixed-means Gaussian hidden Markov model

In this section, the results of applying the Baum-Welch algorithm to learn the FMGHMM parameters θ are presented. The model parameters here do not include the means of the observation distributions $\boldsymbol{\mu}$ since they were predefined based on the model proposed in Ref. 22 as described in Sec. II C.

Table III presents the results of the FMGHMM in regard to the performance metrics. The K-S test score K was 0.118 for Hawaii for both the training and test datasets. For Norrköping, the K scores were 0.047 and 0.045, respectively, for the training and test datasets. In regard to the MAE of the ACF ϵ, it was high, 0.22 or 22%, for both locations. The training time was 6 and 5 s for Hawaii and Norrköping, respectively.

In regard to the observed distribution, Fig. 7 presents a comparison between the test data and samples from the FMGHMM. The generated PDF seems—by visual inspection—to adequately represent the PDF of the CSI. Nonetheless, the accuracy of the model was higher in the case of Norrköping in comparison with Hawaii in regard to the observed distribution, which was expected from the results of the K-S test scores presented before.

Table IV shows the parameters of the fitted distributions. Note that the mean of the obscured state μ_1 was estimated for both locations, using Eq. (9), to be 0.351 and 0.307 for Hawaii and Norrköping, respectively. However, the peaks of the first mode of the CSI PDF were at 0.333 and 0.19 for the Hawaii and Norrköping test datasets, respectively.

Another important observation is that the Baum-Welch algorithm did not fit the variances as perhaps expected. State 2, the state of unobscured with clear-sky, is expected to have a narrow distribution, i.e., small variance, due to the concentration of the observed CSI around 1.0 when the sky is clear. On the other hand, the distribution

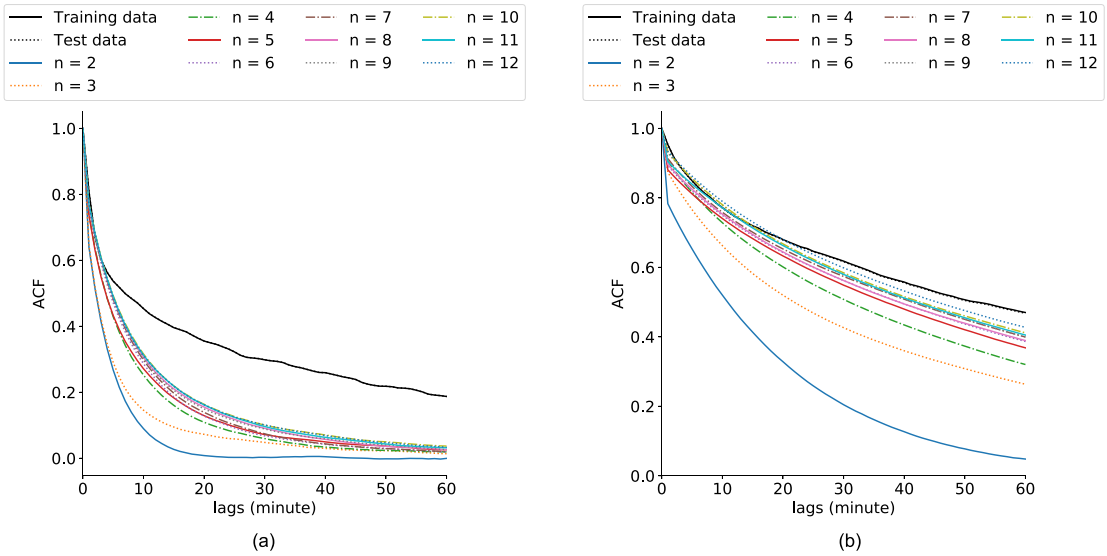

FIG. 5. The ACF of the training and test data and samples from the GHMMs with different numbers of hidden states n. (a) Hawaii and (b) Norrköping. In both locations, the training and test data had the same ACF.

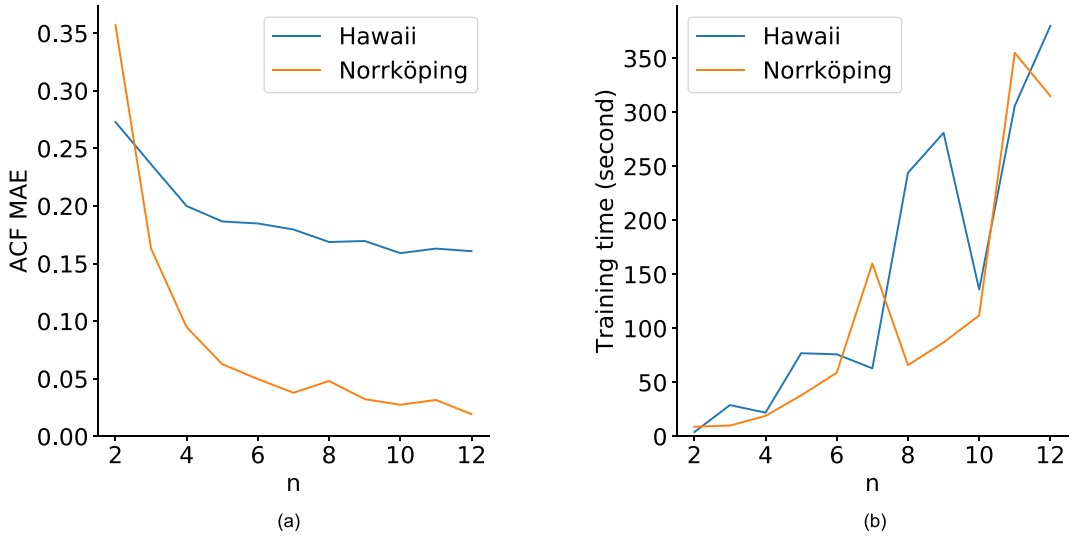

FIG. 6. (a) The MAE of the ACF, ϵ, see Sec. II D, and (b) the training time for the GHMMs with different numbers of hidden states n.

of state 3, unobscured with partially cloudy sky, is expected to be wider than that of state 2. As shown in Table IV, the Baum-Welch algorithm found that the optimal fitted models have a wider distribution in state 2 than in state 3. It should be noted that the Baum-Welch algorithm finds the model with highest log-likelihood $\mathcal{L}$.

Figure 8 shows the ACF of the training and test data along with samples from FMGHMMs for both locations for up to 60 min lags. The ACFs of the training and test data were similar. This was true for both locations. The models produced exponentially decaying ACFs as expected. However, there was a discrepancy between the ACFs of the model and that of the data in both locations. The MAE of the ACF ϵ was 0.22, or 22%, for both locations, see Table III.

Table III presents the performance metrics of the FMGHMM. As comparison, Table II presents the same performance metrics of the GHMMs with $n = 3$. The log-likelihood $\mathcal{L}$ was lower for both locations using the FMGHMM in comparison with the GHMM. The K-S test score K was higher for Hawaii using the FMGHMM compared to the GHMM, 0.118 compared to 0.05. For Norrköping, the differences were negligibly small, 0.003, between the two models in regard to the K value of the test dataset. The MAE of the ACF ϵ, when using the FMGHMM instead of the GHMM, was lower by 0.02 in the case of Hawaii and higher by 0.06 in the case of Norrköping.

TABLE III. The performance metrics for the FMGHMM for both locations and datasets. Note that the MAE of the ACF ϵ was only estimated on the test dataset. On the other hand, the log-likelihood $\mathcal{L}$ and the K-S test score K were estimated for both datasets.

	Training		Test		
Location	$\mathcal{L}$	K	$\mathcal{L}$	K	ϵ
Hawaii	18883.97	0.118	18897.70	0.118	0.22
Norrköping	24275.97	0.047	24222.60	0.045	0.22

To summarize the comparison between the GHMM and the FMGHMM, the GHMM—when compared to the FMGHMM—was better for Hawaii in the K-S test score and worse in the MAE of the ACF, while for Norrköping, the GHMM was comparable in the K-S test score and better in the MAE of the ACF.

C. Improving the autocorrelation function

In this section, we improve the ACF of the HMMs. The results of this improvement on the FMGHMM are presented.

The ACF of the HMM can be calculated directly using Eq. (11). A careful inspection of Eq. (11) shows that the ACF of the samples depends on both the hidden state trajectory and the observation distributions. The transition matrix A controls how likely the process is to persist in a certain hidden state, thus controlling the hidden state trajectory. The variables $\boldsymbol{\pi}$, $\boldsymbol{\mu}$, and $\boldsymbol{\sigma}^2$ control the impacts of the dispersion of the observation distributions on the ACF.

For example, we can take the distributions for Hawaii presented in Table I and assume an HMM with a hidden state trajectory $\{1, 1, \ldots, 2, 2, \ldots, 3, 3, \ldots\}$, i.e., a highly persistent hidden Markov process with $\boldsymbol{\pi} = [1/3, 1/3, 1/3]$. The ACF of such a process will highly depend on the dispersion of the observation distributions, i.e., $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$. This is a result of the fact that the consecutive samples from the process are conditionally independent given the states, more formally $(X_t \perp X_{t-1} | S_t)$. This is to say that consecutive samples are independent even if they were drawn from the same distribution—or state, see Fig. 1.

On the other hand, an HMM with the same stationary distribution $\boldsymbol{\pi}$ and observation distributions as in the previous example but with a randomly shuffled hidden state trajectory will have a lower ACF for the same lags as the previous example. This is due to both the independent sampling of the observations and the rapidly switching hidden state trajectory. A comparison between the ACFs of the two example processes with 9000 time-steps is provided in Fig. 9.

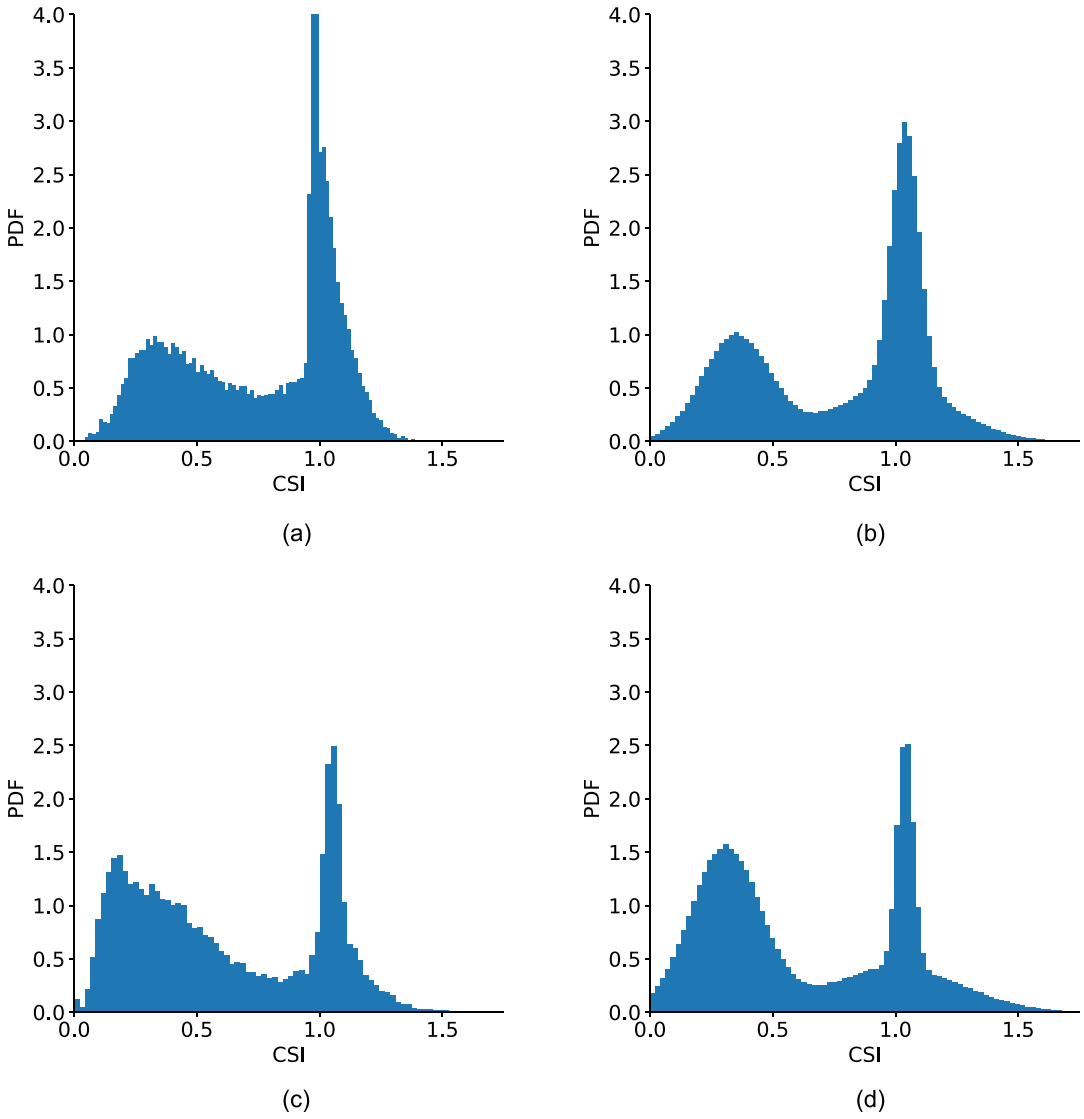

FIG. 7. Histograms of the CSI of the test data (a) and (c); samples from the FMGHMMs (b) and (d). (a) and (b) present the results of Hawaii. (c) and (d) present the results of Norrköping.

Any adjustments to the variables $\boldsymbol{\pi}$, $\boldsymbol{\mu}$, and $\boldsymbol{\sigma}^2$ of the model change the distribution of the CSI generated by the model. Consequently, the only method to improve the ACF of the HMM is by adjusting the transition matrix A while maintaining the stationary distribution $\boldsymbol{\pi}$ constant.

TABLE IV. The observation distributions for the FMGHMM for both locations. All distributions are Gaussians.

	Hawaii		Norrköping	
State	μ_i	σ_i^2	μ_i	σ_i^2
1	0.351	0.0198	0.307	0.0204
2	1.0	0.0526	1.0	0.0692
3	1.04	0.0031	1.04	0.0010

As proven in the Appendix, a scaling constant $\phi \in (0, 1)$ can be multiplied by the off diagonal elements of A followed by rescaling of the diagonal elements using $p_{ii} = 1 - \phi \sum_{j \neq i} p_{ij}$. Such a method, first, will ensure that A satisfies Eq. (3), second, it will not change the stationary distribution $\boldsymbol{\pi}$ of the process, and third, it will alter the ACF by making the hidden state trajectory more persistent.

Here, the authors scaled the previously fitted FMGHMMs in Sec. III B to improve the ACF. The values of the scaling constant ϕ were chosen arbitrarily in the set $\{1, 0.8, 0.6, 0.5, 0.4, 0.3, 0.2\}$. Note that $\phi = 1$ represents the same models fitted in Sec. III B, i.e., no scaling takes place.

Table V compares the scaled models with regard to the log-likelihood $\mathcal{L}$, the K-S test score K, and the MAE of the ACF ϵ. As seen in the table, $\mathcal{L}$ was the highest for the previously fitted models with no scaling, i.e., $\phi = 1$. This explains why the Baum-Welch algorithm

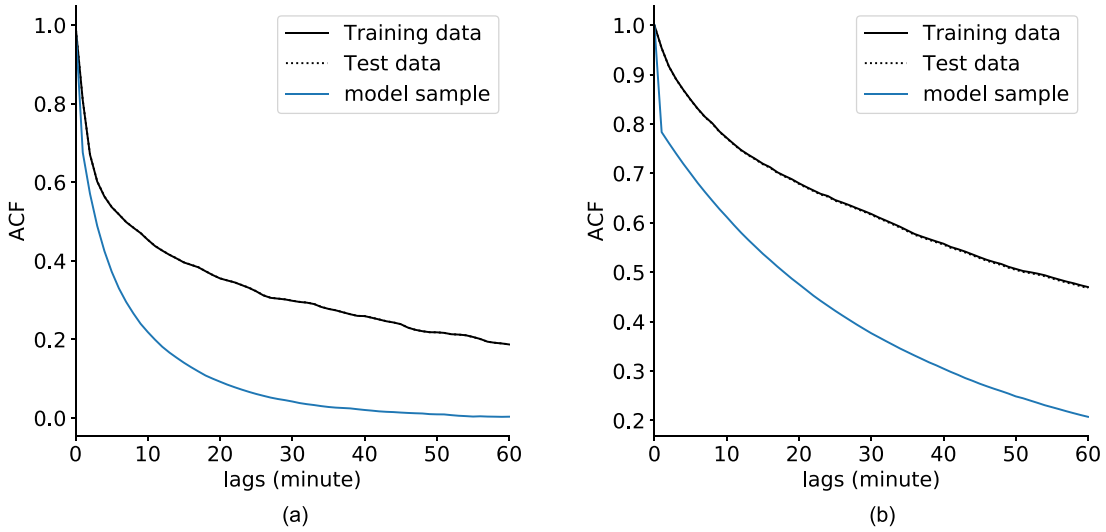

FIG. 8. The ACF for the fitted FMGHMMs, for Hawaii in (a) and Norrköping in (b). In both locations, the training and test data had the same ACF.

previously selected these models. The Baum-Welch algorithm blindly maximizes the log-likelihood with little regard to the remaining performance metrics.[45]

Further research is needed to explain why improving the ACF deteriorated the log-likelihood in our case. The authors suspect that the intersecting distributions of states 2 and 3 might explain this phenomenon. The distribution of state 3 is narrower than that of state 2, see Table IV, which might cause the model to have higher likelihood by switching to state 3 in comparison with persisting in state 2. Nonetheless, further research investigating this behavior is needed, as stated before.

Table V also shows that the differences in the K-S test score were negligibly small. Thus, the proposed scaling did not impact the generated CSI distribution, also see Fig. 10. On the other hand, the proposed scaling improved the MAE of the ACF ϵ compared to the original—unscaled—models. Figure 11 presents the ACFs of samples from the scaled models. A significant improvement in the ACFs can be observed.

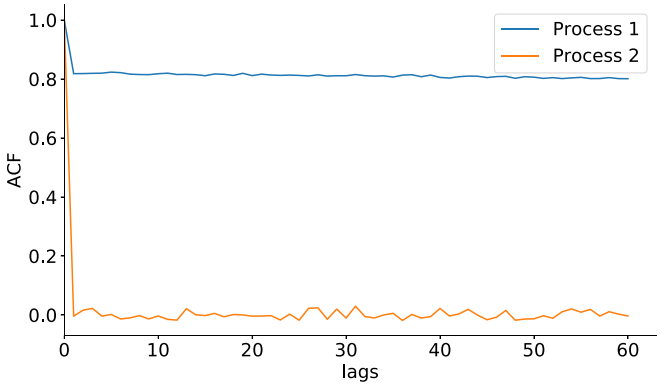

FIG. 9. An example of the ACF of samples from two HMM processes. Both the processes have the same observation distributions as the ones presented for Hawaii in Table I, and both have $\pi = [1/3, 1/3, 1/3]$. Process 1 has a hidden state trajectory $\{1, 1, \ldots, 2, 2, \ldots, 3, 3, \ldots\}$, and process 2 randomly shuffles this hidden state trajectory. In other words, process 2 has an uncorrelated random hidden state trajectory.

D. Comparison with previously proposed models

In this section, a comparison between our proposed models and previous models is provided. We compare our results with those of previous models which used the same datasets and performance metrics.

Table VI compares the models in regard to the K-S test score and the MAE of the ACF. The GHMM did perform comparably with other models when it comes to the K-S test score K. However, when it comes to the MAE of the ACF ϵ, it lagged behind other models. It is important to note that we did not scale the GHMM. If we scale the GHMMs using the method proposed in Sec. III C, the ACF likeness improved. For Hawaii, scaling with $\phi = 0.2$ resulted in the lowest ϵ among the proposed values of ϕ. In this case, we achieved $\epsilon = 0.051$ and $K = 0.050$. Similarly, for Norrköping, scaling with $\phi = 0.4$ achieved $\epsilon = 0.012$ and $K = 0.042$.

Compared to other proposed models, the FMGHMM scaled with $\phi = 0.3$ for Hawaii did not perform well in regard to the K-S test score K, neither did any FMGHMM presented in Table V. For Norrköping, this problem was not observed. Moreover, the MAEs of

TABLE V. The performance metrics of the proposed scaling method tested on the test data and using the FMGHMM. Note that the case $\phi = 1$ is the original fitted models, presented before in Sec. III B.

	Hawaii			Norrköping		
ϕ	$\mathcal{L}$	K	ϵ	$\mathcal{L}$	K	ϵ
1	18897.70	0.118	0.22	24222.60	0.045	0.22
0.8	18820.36	0.117	0.19	24205.54	0.045	0.17
0.6	18541.54	0.116	0.15	24142.82	0.047	0.10
0.5	18288.97	0.114	0.12	24084.92	0.044	0.08
0.4	17922.33	0.115	0.09	23999.54	0.048	0.03
0.3	17383.42	0.117	0.05	23871.73	0.045	0.03
0.2	16547.13	0.112	0.06	23668.90	0.048	0.07

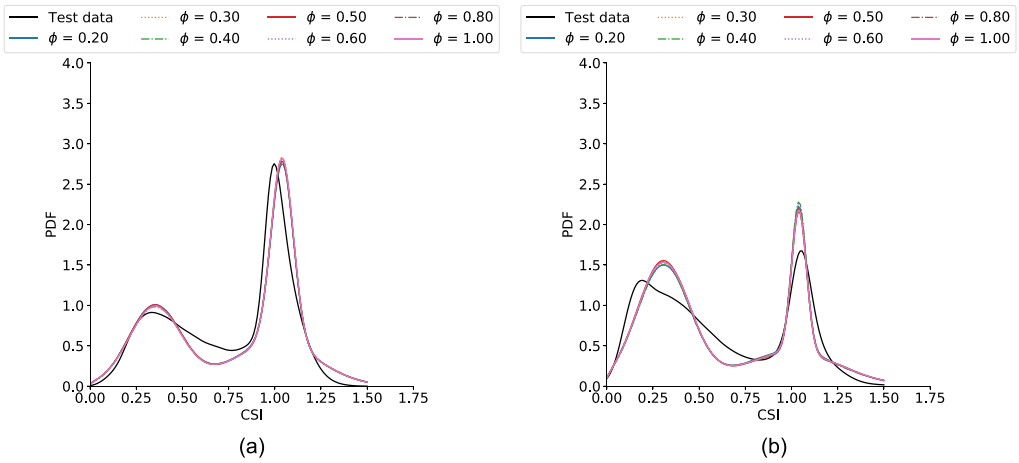

FIG. 10. Distribution of the test data and samples from the FMGHMMs with different scaling constants ϕ. (a) Hawaii and (b) Norrköping.

the ACF of the FMGHMM in both locations were close to those of previous models.

IV. DISCUSSION

This paper extends the previous models in Refs. 5 and 18 by considering a Gaussian observation distribution. In addition, it connects the new model to previous meteorological CSI models and proposes a method to improve the ACF likeness between samples from the fitted models and CSI data.

Two HMMs were developed: GHMM and FMGHMM. In the GHMM, the model parameters were fully learned from the data. The FMGHMM predefined the means of the observation distribution of each hidden state. Still, the remaining model parameters were learned from the data.

In the GHMM, we varied the number of hidden states from 2 to 12. Since the model parameters were freely learned from the data, the fitted distributions for $n = 3$ were not guaranteed to resemble the cloud cover categories proposed by Ref. 22, see Table I.

The K-S test score K of the GHMMs improved in both locations as the number of hidden states increased, see Figs. 3(c) and 3(d). The MAE of the ACF ϵ also improved as the number of hidden states increased; however, it quickly stagnated for Hawaii, see Fig. 6(a).

In regard to the FMGHMM, three hidden states were defined: obscured, unobscured in clear sky, and unobscured in partially cloudy sky. These states were defined following Hollands and Suehrcke,[22] and they represent the cloud cover above a certain location. The means of the Gaussian observation distributions of the three hidden states were predefined using the fraction of bright sunshine. The Baum-Welch algorithm was then used to learn the remaining model parameters.

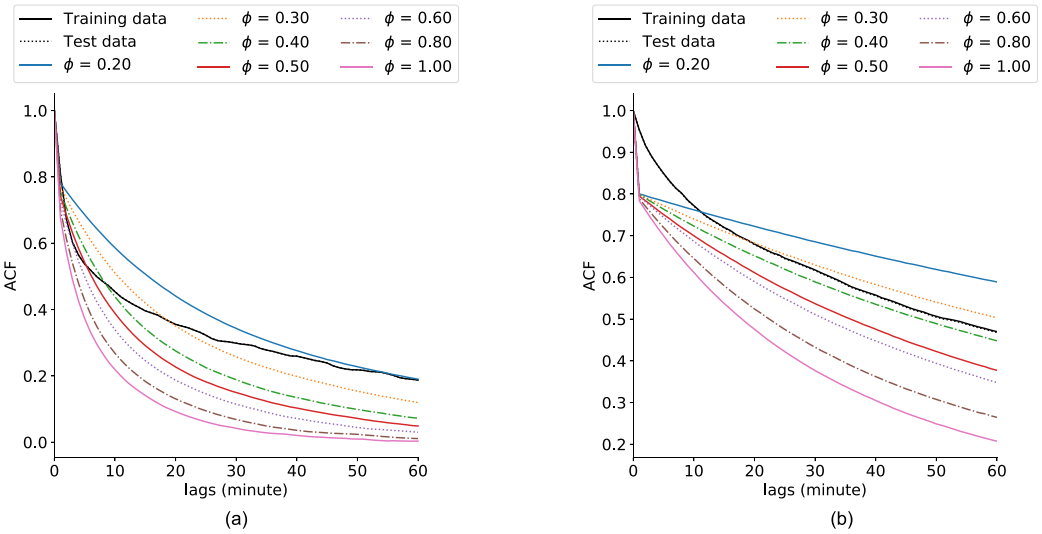

FIG. 11. The ACF of the test data and samples from the FMGHMMs with different scaling constants ϕ. (a) Hawaii and (b) Norrköping. In both locations, the training and test data had the same ACF.

TABLE VI. Comparison of the performance of the proposed models with the previously developed models that used the same datasets—the results of the previously developed models were presented before in Ref. 5.

	Hawaii		Norrköping	
Model	K	ϵ	K	ϵ
References 17 and 26	0.004	0.028	0.009	0.053
Reference 18	0.053	0.077	0.077	0.087
Reference 5, $n = 30$	0.047	0.038	0.044	0.023
GHMM, $n = 3$	0.050	0.240	0.048	0.160
FMGHMM, scaled $\phi = 0.3$	0.117	0.050	0.045	0.030

FMGHMMs produced K-S test scores, K, of 0.118 for Hawaii and 0.045 for Norrköping. Unexpectedly, the variances of the fitted observation distributions were wider for the unobscured in the clear-sky state than for the unobscured in the partially cloudy state, see Table IV.

Comparing the GHMM with $n = 3$ and FMGHMM showed that the K-S test score K increased by 0.068 for Hawaii and decreased by 0.003 for Norrköping (negligible decrease). The MAE of the ACF ϵ decreased by 0.02 for Hawaii and increased by 0.06 for Norrköping.

Increasing the number of hidden states improved the GHMM K-S test score. This, nonetheless, comes at the cost of the training time. Moreover, the K-S test scores of the simpler GHMMs, the models with fewer hidden states, were still comparable with the previously developed models in the literature, e.g., Refs. 5, 17, 18, and 26.

Fitting the HMM using the Baum-Welch algorithm might cause the ACF of the HMM to underperform when compared to the training data. A model with slightly lower log-likelihood might have higher ACF likeness compared to the Baum-Welch optimal model. In this paper, a method to improve the ACF likeness of the HMM models is proposed. This method ensures that the generated distribution from the HMM does not change.

Generating CSI time-series from a fitted model is a two-step process. In the first step, the hidden state trajectory, i.e., the Markov process $\{S_1, \ldots, S_T\}$, is sampled using the method detailed in Ref. 58. In the second step, the CSI observations can be randomly sampled from the distribution associated with each hidden state generated in the first step.

V. CONCLUSIONS

This paper showed that hidden Markov models (HMMs) with Gaussian observation distributions can be employed as generative models for the clear-sky index (CSI). These models can also be adapted to predefine the means of the Gaussian distribution associated with each hidden state. Fitting the models solely by maximizing the log-likelihood might cause the models to underperform when it comes to other performance metrics, e.g., autocorrelation function (ACF) likeness. Care must be taken to ensure that the ACF of the fitted HMM resembles the ACF of the CSI data. In this paper, a novel method to improve the ACF likeness of HMMs was proposed. Increasing the number of hidden states can improve the goodness-of-fit of the CSI distribution. This, however, comes at the cost of the training time, which grows as a function of the square of the number of hidden states.

Future contributions might expand this work by employing other distributions for the hidden states, e.g., lognormal and Weibull distributions. In addition, future works might explore the impacts of using autoregressive HMMs on the ACF likeness of the CSI model. Testing the performance of the method on various CSI time-series and in different conditions, e.g., solar altitude angles and air masses, is encouraged for future works. Finally, developing a method, similar to the one proposed here, to generate the direct component of the irradiance is left for future works.

ACKNOWLEDGMENTS

This work was financially supported by SamspEL 2016–2020 in the project "Development and evaluation of forecasting models for solar power and electricity use over space and time," financed primarily by the Swedish Energy Agency. This work forms part of the Swedish strategic research programme StandUp for Energy.

APPENDIX: IMPROVING THE ACF OF THE HMM

As stated in Eq. (5), the stationary distribution $\boldsymbol{\pi}$ of the Markov chain can be estimated by solving

$$\begin{aligned} \boldsymbol{\pi} A &= \boldsymbol{\pi}, \\ \boldsymbol{\pi}(A - I) &= \mathbf{0}, \end{aligned} \tag{A1}$$

where I is the identity matrix. Notice that multiplying both sides of Eq. (A1) by any constant ϕ does not change the solution. Equation (A1) can be expanded by substituting the diagonals of A by $p_{ii} = 1 - \sum_{j \neq i} p_{ij}$ to

$$[\pi_1, \ldots, \pi_n] \begin{bmatrix} -\sum_{j \neq 1} p_{1j} & \cdots & p_{1n} \\ \vdots & \ddots & \vdots \\ p_{n1} & \cdots & -\sum_{j \neq n} p_{nj} \end{bmatrix} = \mathbf{0}. \tag{A2}$$

Equation (A2) indicates that any constant ϕ will only scale the off diagonal elements of the transition matrix A.

The values of the constant ϕ need to satisfy some conditions. First, ϕ cannot be 0; otherwise; no single solution to Eq. (A1) exists. Second, all the off diagonal elements and the diagonal elements, rescaled as $p_{ii} = 1 - \phi \sum_{j \neq i} p_{ij}$, of the transition matrix A, have to be in [0, 1] as they are probabilities.

Choosing $\phi \in (0, 1)$ will decrease the off diagonal transition probabilities and increase the diagonal transition probabilities of the transition matrix A. Consequently, the process will be more persistent and still has the same stationary distribution $\boldsymbol{\pi}$.

To further explain the effect of the scaling method, note that the largest eigenvalue of the transition matrix A is always 1. The stationary distribution $\boldsymbol{\pi}$ is the left eigenvector corresponding to this eigenvalue, see Eq. (A1) and (Ref. 59, p. 15). Thus, the previously described scaling procedure thus changes the remaining eigenvalues of the transition matrix A. Consequently, the scaling procedure changes the speed of convergence of the Markov chain. The speed of convergence of a Markov chain can be shown to follow (Ref. 59, p. 17)

$$|A^k - A^\infty| \leq \alpha |\lambda_2|^k, \tag{A3}$$

where α is a constant, which satisfies $\alpha > 0$, and λ_i is the an eigenvalue of the matrix A such that $1 = |\lambda_1| \geq |\lambda_2| \geq, \ldots, \geq |\lambda_n|$.

REFERENCES

[1]M. Lave, J. Kleissl, and E. Arias-Castro, "High-frequency irradiance fluctuations and geographic smoothing," Sol. Energy **86**, 2190–2199 (2012).
[2]M. H. Bollen and F. Hassan, *Integration of Distributed Generation in the Power System* (John Wiley & Sons, 2011).
[3]J. Kleissl, *Solar Energy Forecasting and Resource Assessment* (Academic Press, 2013).
[4]M. Lave, M. J. Reno, and R. J. Broderick, "Characterizing local high-frequency solar variability and its impact to distribution studies," Sol. Energy **118**, 327–337 (2015).
[5]J. Munkhammar and J. Widén, "An N-state Markov-chain mixture distribution model of the clear-sky index," Sol. Energy **173**, 487–495 (2018).
[6]J. Deboever, S. Grijalva, M. J. Reno, and R. J. Broderick, "Fast quasi-static time-series (QSTS) for yearlong PV impact studies using vector quantization," Sol. Energy **159**, 538–547 (2018).
[7]C. F. Peruchena, M. Larrañeta, M. Blanco, and A. Bernardos, "High frequency generation of coupled GHI and DNI based on clustered dynamic paths," Sol. Energy **159**, 453–457 (2018).
[8]W. Zhang, W. Kleiber, A. R. Florita, B.-M. Hodge, and B. Mather, "A stochastic downscaling approach for generating high-frequency solar irradiance scenarios," Sol. Energy **176**, 370–379 (2018).
[9]J. Widn and J. Munkhammar, "Spatio-temporal downscaling of hourly solar irradiance data using Gaussian copulas," in 46th IEEE Photovoltaic Specialists Conference (PVSC) (2019).
[10]J. M. Bright, "The impact of globally diverse GHI training data: Evaluation through application of a simple Markov chain downscaling methodology," Renewable Sustainable Energy **11**, 023703 (2019).
[11]G. Lohmann, "Irradiance variability quantification and small-scale averaging in space and time: A short review," Atmosphere **9**, 264 (2018).
[12]J. M. Bright, C. J. Smith, P. G. Taylor, and R. Crook, "Stochastic generation of synthetic minutely irradiance time series derived from mean hourly weather observation data," Sol. Energy **115**, 229–242 (2015).
[13]G. H. Yordanov, T. O. Saetre, and O.-M. Midtgård, "100-millisecond resolution for accurate overirradiance measurements," IEEE J. Photovoltaics **3**, 1354–1360 (2013).
[14]G. M. Lohmann and A. H. Monahan, "Effects of temporal averaging on short-term irradiance variability under mixed sky conditions," Atmos. Meas. Tech. **11**, 3131–3144 (2018).
[15]N. A. Engerer, J. M. Bright, and S. Killinger, "Himawari-8 enabled real-time distributed PV simulations for distribution networks," in 44th IEEE Photovoltaic Specialist Conference (PVSC) (2017), pp. 1405–1410.
[16]Meteotest, *Meteonorm Irradiance Software* (2017).
[17]J. Munkhammar and J. Widén, "An autocorrelation-based copula model for generating realistic clear-sky index time-series," Sol. Energy **158**, 9–19 (2017).
[18]J. Munkhammar and J. Widén, "A Markov-chain probability distribution mixture approach to the clear-sky index," Sol. Energy **170**, 174–183 (2018).
[19]B. J. Brinkworth, "Autocorrelation and stochastic modelling of insolation sequences," Sol. Energy **19**, 343–347 (1977).
[20]A. Skartveit and J. A. Olseth, "The probability density and autocorrelation of short-term global and beam irradiance," Sol. Energy **49**, 477–487 (1992).
[21]J. Munkhammar, J. Rydén, J. Widén, and D. Lingfors, "Simulating dispersed photovoltaic power generation using a bimodal mixture model of the clear-sky index," in 31st European Photovoltaic Solar Energy Conference and Exhibition (EU PVSEC) (2015), pp. 1560–1567.
[22]K. T. Hollands and H. Suehrcke, "A three-state model for the probability distribution of instantaneous solar radiation, with applications," Sol. Energy **96**, 103–112 (2013).
[23]J. Widén, M. Shepero, and J. Munkhammar, "On the properties of aggregate clear-sky index distributions and an improved model for spatially correlated instantaneous solar irradiance," Sol. Energy **157**, 566–580 (2017).
[24]Å. Frimane, T. Soubdhan, J. M. Bright, and M. Aggour, "Nonparametric Bayesian-based recognition of solar irradiance conditions: Application to the generation of high temporal resolution synthetic solar irradiance data," Sol. Energy **182**, 462–479 (2019).
[25]J. M. Bright, O. Babacan, J. Kleissl, P. G. Taylor, and R. Crook, "A synthetic, spatially decorrelating solar irradiance generator and application to a LV grid model with high PV penetration," Sol. Energy **147**, 83–98 (2017).
[26]J. Munkhammar and J. Widén, "An autocorrelation-based copula model for producing realistic clear-sky index and photovoltaic power generation time-series," in 44th IEEE Photovoltaic Specialist Conference (PVSC) (2017), pp. 3067–3072.
[27]A. P. Grantham, P. J. Pudney, and J. W. Boland, "Generating synthetic sequences of global horizontal irradiation," Sol. Energy **162**, 500–509 (2018).
[28]R. Aguiar and M. Collares-Pereira, "TAG: A time-dependent, autoregressive, Gaussian model for generating synthetic hourly radiation," Sol. Energy **49**, 167–174 (1992).
[29]C. Voyant, M. Muselli, C. Paoli, and M.-L. Nivet, "Optimization of an artificial neural network dedicated to the multivariate forecasting of daily global radiation," Energy **36**, 348–359 (2011).
[30]G. M. Lohmann, A. Hammer, A. H. Monahan, T. Schmidt, and D. Heinemann, "Simulating clear-sky index increment correlations under mixed sky conditions using a fractal cloud model," Sol. Energy **150**, 255–264 (2017).
[31]H. Morf, "The stochastic two-state solar irradiance model (STSIM)," Sol. Energy **62**, 101–112 (1998).
[32]R. J. Aguiar, M. Collares-Pereira, and J. P. Conde, "Simple procedure for generating sequences of daily radiation values using a library of Markov transition matrices," Sol. Energy **40**, 269–279 (1988).
[33]E. Palomo, "Hourly solar radiation time series as first-order Markov chains," in Actes du International Solar Energy Society Solar World Congress (1989), pp. 2146–2150.
[34]B. O. Ngoko, H. Sugihara, and T. Funaki, "Synthetic generation of high temporal resolution solar radiation data using Markov models," Sol. Energy **103**, 160–170 (2014).
[35]J. Wegener, M. Lave, J. Luoma, and J. Kleissl, *Temporal Downscaling of Irradiance Data via Hidden Markov Models on Wavelet Coefficients: Application to California Solar Initiative Data* (UC San Diego, 2012).
[36]A. P. Grantham, P. J. Pudney, L. A. Ward, M. Belusko, and J. W. Boland, "Generating synthetic five-minute solar irradiance values from hourly observations," Sol. Energy **147**, 209–221 (2017).
[37]L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. IEEE **77**, 257–286 (1989).
[38]D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 2nd ed. (Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2009).
[39]I. L. MacDonald and W. Zucchini, *Hidden Markov and Other Models for Discrete-Valued Time Series* (CRC Press, 1997), Vol. 110.
[40]S. Frühwirth-Schnatter, *Finite Mixture and Markov Switching Models* (Springer Science & Business Media, 2006).
[41]H. Suehrcke, "On the relationship between duration of sunshine and solar radiation on the Earth's surface: Ångströms equation revisited," Sol. Energy **68**, 417–425 (2000).
[42]H. Suehrcke, R. S. Bowden, and K. T. Hollands, "Relationship between sunshine duration and solar radiation," Sol. Energy **92**, 160–171 (2013).
[43]P. M. Broersen, *Automatic Autocorrelation and Spectral Analysis* (Springer Science & Business Media, 2006).
[44]R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice* (OTexts, Melbourne, Australia, 2018).
[45]T. Rydén, T. Teräsvirta, and S. Åsbrink, "Stylized facts of daily return series and the hidden Markov model," J. Appl. Econometrics **13**, 217–244 (1998).
[46]M. Jurado, J. M. Caridad, and V. Ruiz, "Statistical distribution of the clearness index with radiation data integrated over five minute intervals," Sol. Energy **55**, 469–473 (1995).
[47]M. J. Beal, Z. Ghahramani, and C. E. Rasmussen, "The infinite hidden Markov model," in Advances in Neural Information Processing Systems (2002), pp. 577–584.
[48]See https://github.com/hmmlearn/hmmlearn/ for information about the hidden Markov models package in Python.
[49]L. Theis, A. V. D. Oord, and M. Bethge, "A note on the evaluation of generative models," in International Conference on Learning Representations (ICLR 2016) (2016), pp. 1–10.
[50]F. J. Massey, Jr., "The Kolmogorov-Smirnov test for goodness of fit," J. Am. Stat. Assoc. **46**, 68–78 (1951).
[51]SMHI (SMHI, Norrköping, Sweden, 2008).

[52]M. Sengupta and A. Andreas, "Oahu solar measurement grid (1-year archive): 1-second solar irradiance; Oahu, Hawaii (data)," Technical Report No. NREL/DA-5500-56506 (National Renewable Energy Lab. (NREL), Golden, CO, USA, 2010).
[53]Kipp & Zonen, *CM 21 Precision Pyranometer Instruction Manual* (2019).
[54]LI-COR, *LI-200R Pyranometer* (2019).
[55]P. Ineichen and R. Perez, "A new airmass independent formulation for the Linke turbidity coefficient," Sol. Energy **73**, 151–157 (2002).
[56]Soda-service, *CAMS McClear Service for Estimating Irradiation Under Clear-Sky* (2016).
[57]X. Sun, J. M. Bright, C. A. Gueymard, B. Acord, P. Wang, and N. A. Engerer, "Worldwide performance assessment of 75 global clear-sky irradiance models using principal component analysis," Renewable Sustainable Energy Rev. **111**, 550–570 (2019).
[58]J. Widén, A. M. Nilsson, and E. Wäckelgård, "A combined Markov-chain and bottom-up approach to modelling of domestic lighting demand," Energy Build. **41**, 1001–1012 (2009).
[59]P. Lorek, "Speed of convergence to stationarity for stochastically monotone Markov chains," Ph.D. thesis (University of Wrocław, 2007).