Solar Energy 173 (2018) 487–495

Contents lists available at ScienceDirect

Solar Energy

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

An *N*-state Markov-chain mixture distribution model of the clear-sky index

Joakim Munkhammar*, Joakim Widén

Built Environment Energy Systems Group, Department of Engineering Sciences, Uppsala University, SE-751 21 Uppsala, Sweden

ARTICLE INFO

Keywords:
Clear-sky index
Markov-chain modeling
Mixture distribution modeling
Autocorrelation function

ABSTRACT

This paper presents an *N*-state Markov-chain mixture distribution approach to model the clear-sky index. The model is based on dividing the clear-sky index data into bins of magnitude and determining probabilities for transitions between bins. These transition probabilities are then used to define a Markov-chain, which in turn is connected to a mixture distribution of uniform distributions. When trained on measured data, this model is used to generate synthetic data as output. The model is an *N*-state generalization of a previously published two-state Markov-chain mixture distribution model applied to the clear-sky index. The model is tested on clear-sky index data sets for two different climatic regions: Norrköping, Sweden, and Oahu, Hawaii, USA. The model is also compared with the two-state model and a copula model for generating synthetic clear-sky index time-series as well as other existing clear-sky index generators in the literature. Results show that the *N*-state model is generally on par with, or superior to, state-of-the-art synthetic clear-sky index generators in terms of Kolmogorov–Smirnov test statistic, autocorrelation and computational speed.

1. Introduction

There is a challenge in quantifying, and reproducing, solar irradiance variability over time. Normalized solar irradiance, expressed as the clear-sky index (CSI), has interesting statistical features, particularly on minute to instantaneous scale (Bright et al., 2015). The distribution of the clear-sky index is typically characterized by two or three peaks, consequently modeling is often focused on mixture distributions to construct realistic distributions (Hollands and Huget, 1983; Suehrcke and McCormick, 1988; Hollands and Suehrcke, 2013; Munkhammar et al., 2015; Widén et al., 2017).

Not only the probability distribution of the clear-sky index is of interest in modeling the clear-sky index, but also the accuracy of temporal variability is important (Bright et al., 2017). Time-series realism is typically measured via autocorrelation function similarity over a set of lags (Munkhammar and Widén, 2017a; Brinkworth, 1977; Skartveit and Olseth, 1992). The autocorrelation function of the clear-sky index time-series has been studied previously for instantaneous solar irradiance (Munkhammar and Widén, 2017a; Brinkworth, 1977; Skartveit and Olseth, 1992; Aguiar and Collares-Pereira, 1992; Hansen et al., 2010; Munkhammar and Widén, 2018; Munkhammar and Widén, 2017b), where the autocorrelation function has been shown to be positive and follow an exponential slope for hour resolution (Skartveit and Olseth, 1992; Aguiar and Collares-Pereira, 1992; Hammer and Beyer, 2013), while it has also been shown to have negative values for minute resolution (Munkhammar and Widén, 2017a; Hansen et al., 2010; Munkhammar and Widén, 2017b).

Models aiming to quantify the clear-sky index and generating realistic synthetic clear-sky index data, including proper autocorrelation function similarity, include Gaussian-Markov (Brinkworth, 1977), autoregressive Gaussian (Aguiar and Collares-Pereira, 1992), neural networks (Voyant et al., 2011), copula modeling (Munkhammar and Widén, 2017a; Munkhammar and Widén, 2017b), fractal cloud modeling (Lohmann et al., 2017) and Markov-chains (Bright et al., 2015; Morph, 1998; Aguiar et al., 1988; Palomo, 1989; Ngoko et al., 2014).

These clear-sky or clearness index generators are practical, since they use some existing data set of lower resolution or averaged clear-sky index or clearness data to estimate higher resolution data (temporal or spatial), see e.g. (Bright et al., 2015; Bright et al., 2017; Morph, 1998; Ngoko et al., 2014; Wegener et al., 2012; Grantham et al., 2017), or smaller amounts of clear-sky index data to generate unlimited amounts of data (Munkhammar and Widén, 2017a; Munkhammar and Widén, 2018; Munkhammar and Widén, 2017b). In particular Markov-chains have been useful as clear-sky index generators, where (Ngoko et al., 2014) generated minute resolution, while (Aguiar et al., 1988) generated daily resolution. In Munkhammar and Widén (2018) a two-state Markov-chain mixture probability distribution was used, while in Morph (1998) a general model for generating clear and cloudy periods was constructed. Generally, it should be emphasized that models for generating synthetic clear-sky index data are useful complements to

* Corresponding author.
E-mail addresses: joakim.munkhammar@angstrom.uu.se (J. Munkhammar), joakim.widen@angstrom.uu.se (J. Widén).

https://doi.org/10.1016/j.solener.2018.07.056
Received 13 June 2018; Received in revised form 9 July 2018; Accepted 20 July 2018

irradiation estimates from ground measurements or from satellite data, see e.g. (Engerer et al., 2017), and software, such as Meteonorm (2017).

One conclusion from the literature is that designing a comparatively simple and easily computable model that generates accurate synthetic clear-sky index data is a challenge. Here Morph (1998) is a particular example of a comparatively simple model with complex dynamical output. It was based on two-states, like the two-state Markov-chain mixture model, and it was hypothesized that an N-state model could perhaps be used for obtaining higher time-series similarity (Munkhammar and Widén, 2018).

This study aims to develop an N-state Markov-chain mixture distribution model for the clear-sky index. The model is a direct generalization of the two-state Markov-chain mixture distribution model recently applied to generate synthetic clear-sky index time-series (Munkhammar and Widén, 2018). The model is trained and tested on solar irradiance data sets for Norrköping and Hawaii, which were also used in Munkhammar and Widén (2017a) and Munkhammar and Widén (2018). This model is a clear-sky index generator, similar to the models in Munkhammar and Widén (2017a), Munkhammar and Widén (2017b), Ngoko et al. (2014), Grantham et al. (2017) and Grantham et al. (2018). Although also Markov-chain based, the model in this study requires less input than the model in Ngoko et al. (2014), which uses hourly weather observations including sea level air pressure, wind speed, cloud base height and cloud cover. Also, only global horizontal irradiance (GHI) is used, and not direct normal irradiance (DNI), like in Grantham et al. (2017). The model utilizes clear-sky index data as input for training the model similarly to the copula model (Munkhammar and Widén, 2017a; Munkhammar and Widén, 2017b) and the two-state Markov-mixture distribution model (Munkhammar and Widén, 2018), and produces unbounded outputs of clear-sky index time-series data.

In comparison, however, the model is less mathematically and computationally complex than for example the copula-based model in Munkhammar and Widén (2017a) and Munkhammar and Widén (2017b) or the model in Grantham et al. (2017). While the two-state model in Munkhammar and Widén (2018) also had a method-part of connecting clear and cloudy states to physical observables such as for example duration of bright sunshine, the N-state generalization developed here lacks this physical basis, and is instead essentially a data-driven computational tool.

This paper is organized as follows. In Section 2 the methodology, including model and data, is presented. In Section 3 the results from the model are presented, in Section 4 the results are discussed, and in Section 5 conclusions are drawn.

2. Methodology

2.1. Markov-chain mixture distribution modeling

This study focuses on modeling the indeterministic temporal variability of instantaneous solar irradiance, the clear-sky index, defined as the variability of the solar irradiance when the deterministic variability of the sun's position on the sky dome has been removed.

Formally, the clear-sky index κ is defined as the ratio between the measured global horizontal irradiance (GHI) $G(t)$ and the estimated global horizontal clear-sky irradiance $G_c(t)$ over time t:

$$\kappa_t \equiv \frac{G(t)}{G_c(t)}. \tag{1}$$

The Markov-chan mixture distribution model developed here is based on the concept of a *Markov-chain mixture distribution*, which is arguably a form of *Hidden Markov Model* (HMM), see (Murphy, 2012, p. 312) for more information on HMM. The model, as mentioned, is an N-state generalization of a previously published two-state model (Munkhammar and Widén, 2018). In that study, the clear-sky index was divided into two states: clear and cloudy, where clear was defined as all CSI values above a certain CSI threshold and cloudy for all CSI values below that threshold. Each state was modeled with a probability distribution, and a Markov-chain was used to determine the transitions between the two states, thereby making it a Markov-chain mixture distribution. From this model synthetic time-series of the clear-sky index were generated. See (Munkhammar and Widén, 2018) for more information on that approach. Here follows an N-state generalization of this model.

As in previous studies by the authors (Munkhammar and Widén, 2017a; Munkhammar and Widén, 2018; Munkhammar and Widén, 2017b) the clear-sky index is here modeled as a stochastic variable κ, and the following mathematical construction defines an N-state Markov-chain mixture distribution for a time-series bounded by real-valued limits $[a, b]$, where $a < b$:

$$\kappa = X_1 Y_1 + X_2 Y_2 + \ldots + X_N Y_N \tag{2}$$

where X_i are determined by a Markov-chain and Y_i are independent stochastic variables. Let the Markov-chain be a homogenous temporal Markov-chain with N states, where the outcome for each occupied state $i \in [1, N]$ is defined by an $N \times 1$ vector $X = [X_1, X_2, \ldots, X_N]$ of zeros with the exception of position i which is set to 1, e.g. $X = [0, \ldots, 0, 1, 0, \ldots, 0]$. This makes the outcomes of the Markov-chain a ***Multinoulli distribution***, a categorical distribution, which in general is a stochastic variable that can take on any of N number of elementary events (Murphy, 2012).

This construction makes each stochastic variable X_i automatically Bernoulli distributed. Furthermore, let the stochastic variables Y_i be uniformly distributed on the interval $[\frac{i-1}{N}(b-a) + a, \frac{i}{N}(b-a) + a[$. The combination of the categorical distribution for X_i and the particular uniform distributions for Y_i in Eq. (2) generates a stochastic variable κ, which sampled for each time-step generates a time-series that samples from the stochastic variable Y_i at each time-step:

$$[Y_1, Y_1, Y_3, Y_4, Y_1, Y_5, \ldots], \tag{3}$$

where the chosen i for each time-step is dependent on the occupied state of the Markov-chain state in that time-step. In terms of probability distribution, this generates a probability density function f:

$$f(\kappa) = \begin{cases} P(x_1 \leqslant \kappa < x_2)\frac{1}{\Delta x}, & x_1 \leqslant \kappa < x_2 \\ P(x_2 \leqslant \kappa < x_3)\frac{1}{\Delta x}, & x_2 \leqslant \kappa < x_3 \\ \vdots \\ P(x_N \leqslant \kappa < x_{N+1})\frac{1}{\Delta x}, & x_N \leqslant \kappa < x_{N+1}, \end{cases} \tag{4}$$

where for $i \in [1, N]$ one can define $x_i = \frac{i-1}{N}(b-a) + a$, $\Delta x = |b-a|/N$ and $P_i = P(x_i \leqslant \kappa < x_i + \Delta x)$, which is equal to the *stationary distribution* or *stable state* of the Markov-chain for state i, see (Cinlar, 1975, p. 243) for information on stable states. $1/\Delta x$ is defined as the probability density distribution of each uniform distribution. By definition of probability densities $f_i(\kappa)$, defined on the disjoint sets $\kappa \in [x_i, x_{i+1}]$, we have the following:

$$F(\kappa) = \sum_{i=1}^{N} f_i(\kappa)\Delta x = \sum_{i=1}^{N} P(x_i \leqslant \kappa < x_i + \Delta x) \tag{5}$$

where $F(x)$ is the cumulative distribution function of κ, defined on $[a, b]$, and equal to unity for $F(b)$. The limit of letting $N \to \infty$, and thereby $\Delta x \to 0$, brings:

$$\lim_{N\to\infty} \sum_{i=1}^{N} P(x_i \leqslant \kappa < x_i + \Delta x) = \int_a^{\kappa} f(x)dx, \tag{6}$$

which equals the continuous limit version of the cumulative distribution function of f. The model, in conceptual and algorithmic terms, is presented in Fig. 1.

The key to using this model is training it on a training data set; a time-series of the clear-sky index. The first step is setting the number of states N, then dividing the data set into N bins $[x_i, x_{i+1}]$ for $i \in [1, N]$.