Statistical Analysis of Multi-day Solar Irradiance using a Threshold Time Series model

Carolina Euán *†, Ying Sun‡ and Brian J Reich §

January 4, 2022

Abstract

The analysis of solar irradiance has important applications in predicting solar energy production from solar power plants. Although the sun provides every day more energy than we need, the variability caused by environmental conditions affects electricity production. Recently, new statistical models have been proposed to provide stochastic simulations of high-resolution data to downscale and forecast solar irradiance measurements. Most of the existing models are linear and highly depend on normality assumptions. However, solar irradiance shows strong non-linearity and is only measured during the day time. Thus, we propose a new multi-day threshold autoregressive (TAR) model to quantify the variability of the daily irradiance time series. We establish the sufficient conditions for our model to be stationary, and we develop an inferential procedure to estimate the model parameters. When we apply our model to study the statistical properties of observed irradiance data in Guadeloupe island group, a French overseas region located in the Southern Caribbean Sea, we are able to characterize two states of the irradiance series. These states represent the clear-sky and non-clear sky regimes. Using our model we are able to simulate irradiance series that behave similarly to the real data in mean and variability, and more accurate forecasts compared to linear models.

Keywords: Time series, Non-linear models, TAR model, Clear-sky index, Weighted least squares.

*Corresponding author: c.euancampos@lancaster.ac.uk

†Department of Mathematics and Statistics, Lancaster University, Lancaster, LA1 4YF, UK

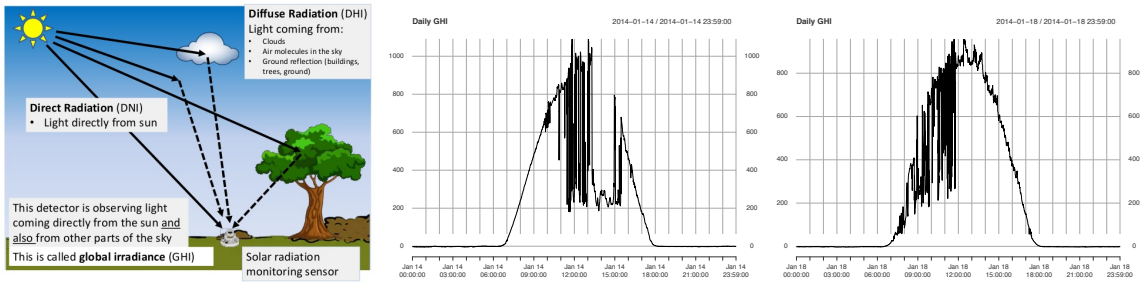
‡Statistics Program, King Abdullah University of Science and Technology, Thuwal 23955-6900, SA

§Department of Statistics, North Carolina State University, Raleigh, NC

1 Introduction

Human activity strongly depends on electricity production. Currently, major sources for generation of electricity, such as fossil fuels (coal and gas), are non-sustainable sources of energy, producing air, water, and land pollution. Therefore, there is a high interest in renewable energy production which grows globally very rapidly. Among different forms of renewable energies, solar photovoltaic (PV) has been the most popular performer for the second year in a row, with a newly installed capacity increasing by approximately 33%, according to REN21 (2020). Every day, the sun provides more energy than we need to meet many of the challenges facing the world. However, the power production from PV plants may not be stable, due to the varying environmental conditions, such as cloud cover and daily temperature. This variability can have severe consequences on PV technologies; for example, sharp changes (lasting a few seconds) can cause local voltage flicker issues, and a more extended time scale change can affect the storage system (Kleissl, 2013). Thus, it is essential to understand such variability for the management of solar power production.

The amount of electricity generated by PV systems depends on the intensity and wavelength of solar radiation available to the PV device. In this paper, we focus on the analysis of the global horizontal irradiance (GHI) that measures the total hemispheric down-welling solar radiation on a horizontal surface (Figure 1(a)). There is a variability (deterministic) in the GHI daily time series that corresponds to the Sun's movement during the day, and that can be precisely predicted. For example, Figure 1(b) shows two examples of daily GHI time series that raise to a positive value around 6:30 am and drop after 6:00 pm (period depending on location and season of the year). The main interest in GHI data is to characterize the variability that could be caused by clouds, temperature or any other atmospheric conditions. Figure 1(b)(left) shows high variability between 11:00 am and 1:00 pm followed by a long period of low irradiance and Figure 1(b)(right) shows a high variability in the



(a) (b)

Figure 1: (a) Solar radiation measurements (Figure from slidesshare.net) (b) GHI measured by a pyranometer every 30 sec. Daily irradiance variability can be very different from day to day and the drops of the irradiance values can be observed in any time of the day.

morning but a very clear sky afternoon with low variability.

One of the main goals for modeling solar irradiance is to quantify the volatility/variability at a specific location. PV integrated systems rely on this accurate variability estimation, which is relevant to the grid manager to efficiently distribute the power or select a new site for new system installation. In practice, estimating the variability means a precise estimation of the probability distribution that generates the data. The most common strategy is to propose a model that generates synthetic irradiance time series which reproduce similar expected values and variability. For example, Zhang et al. (2019) proposed a model based on a non-Gaussian distribution to generate high-frequency solar irradiance scenarios. Recent developments for analyzing GHI data focus on stochastic simulation of high-resolution data and forecasting of spatiotemporal GHI (Zhang et al., 2021).

Although this method performs extremely well, they need to first classify cloudy and clear days, based on a fixed threshold rule of low and high variability observed in the increments of GHI. A more flexible approach where the days are classified by clustering algorithms was recently introduced by Frimane et al. (2019). Under this model, a posterior

inference using Markov chain Monte Carlo algorithm is applied to estimate daily distributions of the solar irradiance. Munkhammar and Widén (2019) proposed an alternative model with a more robust classification procedure between clear and non-clear days. Previous research efforts in this area also include models with covariates or cloud motion models (Boland et al., 2001; Bright et al., 2015, 2017; Fernández-Peruchena et al., 2015). However, the classes vary among these models and the number of classes may be subjectively selected.

Popular statistical models that allowed regime switching between time-dependent data are the hidden Markov model (HMM) and the mixture chain model. The HMM assumes the presence of a hidden random process that determines the changes in the dynamics of the observed data. An HMM with three hidden states can have an interpretation linked to three categories of cloud cover (Shepero et al., 2019). However, the goodness of fit for this model to generate GHI data can be very poor in some cases. More hidden states need to be considered to improve the fitting, but this increases the computational cost. The use of multiple time series (Fox et al., 2014) could improve these models, but this option has not been explored for irradiance time series. Barbič et al. (2004) propose a mixture model to detect the distinct behavior of motion data. For irradiance data, Munkhammar and Widén (2019) investigated these models in the case where observations are from different spatial locations. In this paper, we model irradiance data without an a priori classification. We allow our model to change the state (clear or non-clear sky) based on its immediate past values. By considering a two-state model, our proposal is interpretable in terms of clear and non-clear sky regimes, which can be observed partially across days. Although a model with more than two regimes could result in a better fitting of the data, the resulting model might lack interpretability.

A feasible approach for modeling GHI is to consider a time series model. Time series models are efficient in forecasting. Accurate predictions are relevant for efficient energy

management, where efficiency means a precise balance between electricity production and consumption at any moment. The grid management requires short, medium, and long-term predictions to feed the electricity networks. Some of the desirable characteristics of such time series model are seasonality and non-linearity. Grantham et al. (2018) proposed a time series model that represents the seasonality with as a cosine and sine functions of the main Fourier frequencies observed on the data and an autoregressive component. To increase the flexibility of their model, a non-stationary white noise component is considered. However, the non-stationary assumption may cause inconsistent estimators. Similarly, Al-Awadhi and El-Nashar (2002) also consider a Fourier expansion to model the seasonality of the irradiance data, and a bilinear time series component to model the stochastic component of daily global radiation. However, a priori division of the data set in clear and non-clear days is also required. Voyant et al. (2020) proposed a prediction model based on the periodic autoregressive (PAR) model coupled with a Box-Cox transform. This model considers a set of coefficients per hour that uses all-day information for estimating, which gives flexibility to the model. Das et al. (2021) proposed a cyclostationary model for short term prediction of hourly solar irradiance. Althoug, for high temporal resolution data, this might result in a large number of coefficients for PAR models or non-identifiable cycles due to the high level of noise variation. To avoid these issues, we propose a non-linear time series approach to model solar irradiance data. Within the existing non-linear time series, threshold models have shown to be generally sufficient enough to model cyclical econometric data (Tong and Lim, 1980). In the literature, we can find some examples of econometric models applied to environmental data (Magnus et al., 2011; Samanta et al., 2011; Grillenzoni and Carraro, 2021).

The GHI series can be transformed to clear-sky index, $K(t)$, as

$$K(t) = \frac{I(t)}{I_{CS}}, \tag{1}$$

where the time unit t are seconds, $I(t)$ is the GHI observed at time t and I_{CS} is the horizontal extraterrestrial irradiance that can be computed using an atmospheric model (see Kleissl, 2013). Then, $K(t)$ isolates the variability caused by random environmental conditions. Generally, researchers model the logarithm of clear-sky index to avoid the restriction of being strictly positive. In this paper, we propose the use of threshold autoregressive model on the logarithm of clear-sky index series to forecast and describe solar irradiance data.

Unlike the typical models developed for econometric time series, we treat daily irradiance time series as independent replicates and propose a new multi-day model, where daily irradiance time series share a common cyclical component but with different intra-day variability. Our model identifies clear and non-clear day periods without any prior classification of the data by introducing a threshold effect. We propose an estimation procedure that takes into account the single day variability and all days common information. The estimated variability per day also serves as a measure of weather variability across days. Additionally, we show that our model can forecast the irradiance time series more accurately than the persistence ensemble model (commonly used on energy forecasting), some machine learning algorithms, PAR models, and ARIMA models.

This paper remaining is organized as follows: Section 2 introduces the real irradiance data set from the Guadeloupe island group and presents the preliminaries about the existing non-linear time series model and its theoretical properties. Section 3 presents our proposed model and the statistical inference procedure for the model parameters. Lastly, Section 4 presents a detailed data analysis of the Guadeloupe island group irradiance series

by applying our multi-day model. This section also compares our model with different competitors and proposes an algorithm for a probabilistic forecast.

2 Irradiance Data and TAR Model

2.1 Data description

Here, we study an irradiance time series, $\{I(t)\}$, from Jan 21, 2011, to Dec 31, 2011. Data are collected in Guadeloupe island group, a French overseas region located in the Southern Caribbean Sea (Lat 16.22 - Long -61.53 approximately, see Figure 2). Data were collected every second, and the global horizontal irradiance is reported. Then, we use the horizontal extraterrestrial irradiance reported at this location by the National Solar Radiation Database (NSRDB) (Sengupta et al., 2018) to compute the clear-sky index, $K(t)$.

Figure 3 shows a subset of the $I(t)$ series that corresponds to March 2011 from 7 : 00 am to 5 : 00 pm. On the top of the GHI series, we plot a dashed line that represents the extraterrestrial irradiance, I_{CS}. On one-second data, a clustering of the $I(t)$ series will result either in one big cluster (non-clear days) or several small clusters (different degrees of non-clearness). Therefore, we do not consider a prior classification of the data. We now compute the logarithm of the clear-sky index defined in (1). Figure 4 shows the $\log K(t)$ series corresponding to March, 2011. We observe that March 9, March 18, and March 31 are days on which the irradiance is close to the horizontal extraterrestrial irradiance, whereas March 05, March 14, and March 25 are days on which we observe a high variability on the horizontal irradiance. In general, irradiance profiles per day are different across the year.

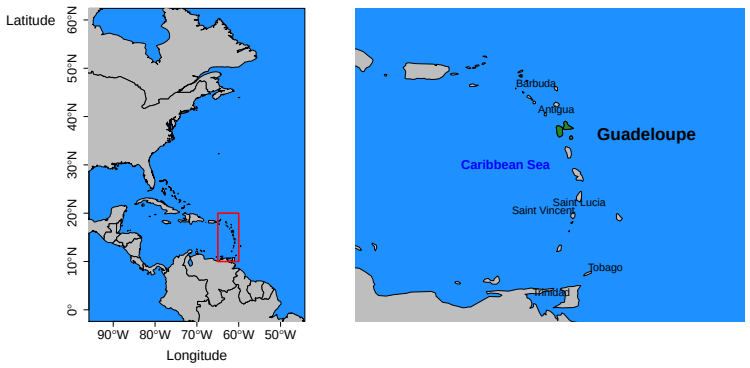


Figure 2: Location of Guadeloupe Island. Right: Large scale map where the island of Guadeloupe is inside the box. Left: the island of Guadeloupe is represented in green. The open sea location encourages the presence of more variable weather conditions.

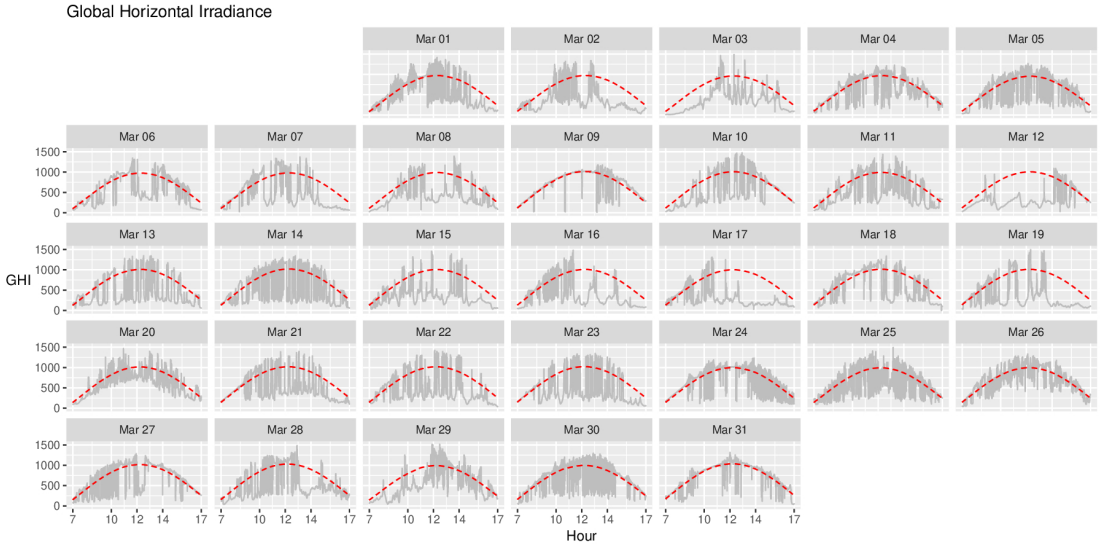


Figure 3: GHI time series for one-month data; the dashed line is the interpolated horizontal extraterrestrial irradiance. Irradiance series can be very different from day to day.

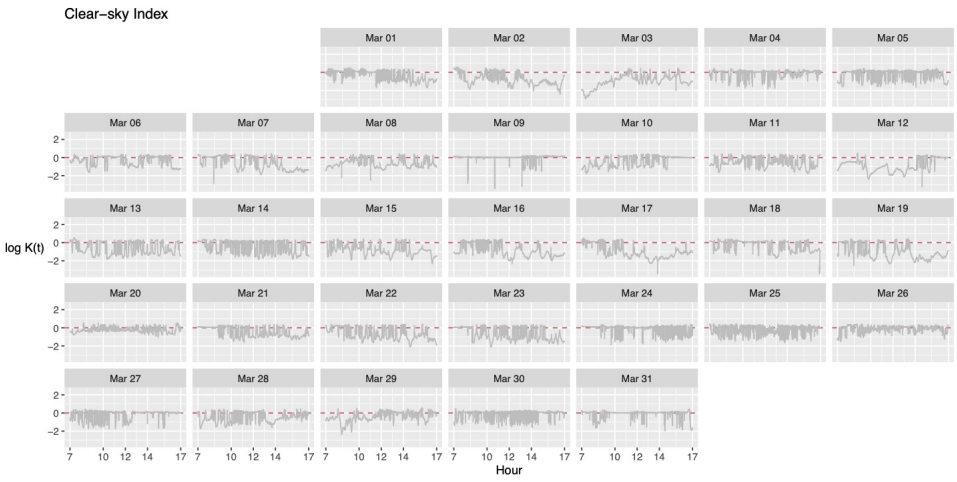


Figure 4: Log $K(t)$ time series for one-month data; the dashed line at 0 represents the irradiance equal to the horizontal extraterrestrial irradiance. Drops on the irradiance value can be observed at any time of the day.

2.2 Nonlinearity and Non-Gaussianity

Linear models are the most commonly used statistical models. However, for complex real data applications, they suffer from certain limitations. For example, under a linear representation of a time series, if a Gaussian white noise is assumed, all the finite-dimensional distributions of the process are expected to be Gaussian. To visualize if the distribution of the $\log K(t)$ series may be Gaussian, we plot the histogram of the series per day in March (see Figure 5). Some days, such as March 14 and March 23, show strong evidence of a bimodal density, which is visually against Gaussianity. The bimodal density strongly suggests a non-gaussian distribution of noise, even more, the feasible presence of regimes. Additionally, we plot the bivariate density contour plots of $(\log K(t), \log K(t+1))$, where we observe empirical evidence of a two-state correlation structure on days like March 13, 14, 21, and 23. These visual tools together are evidence against a linear model.

In the literature, there are several tests to revise whether or not a linear time series

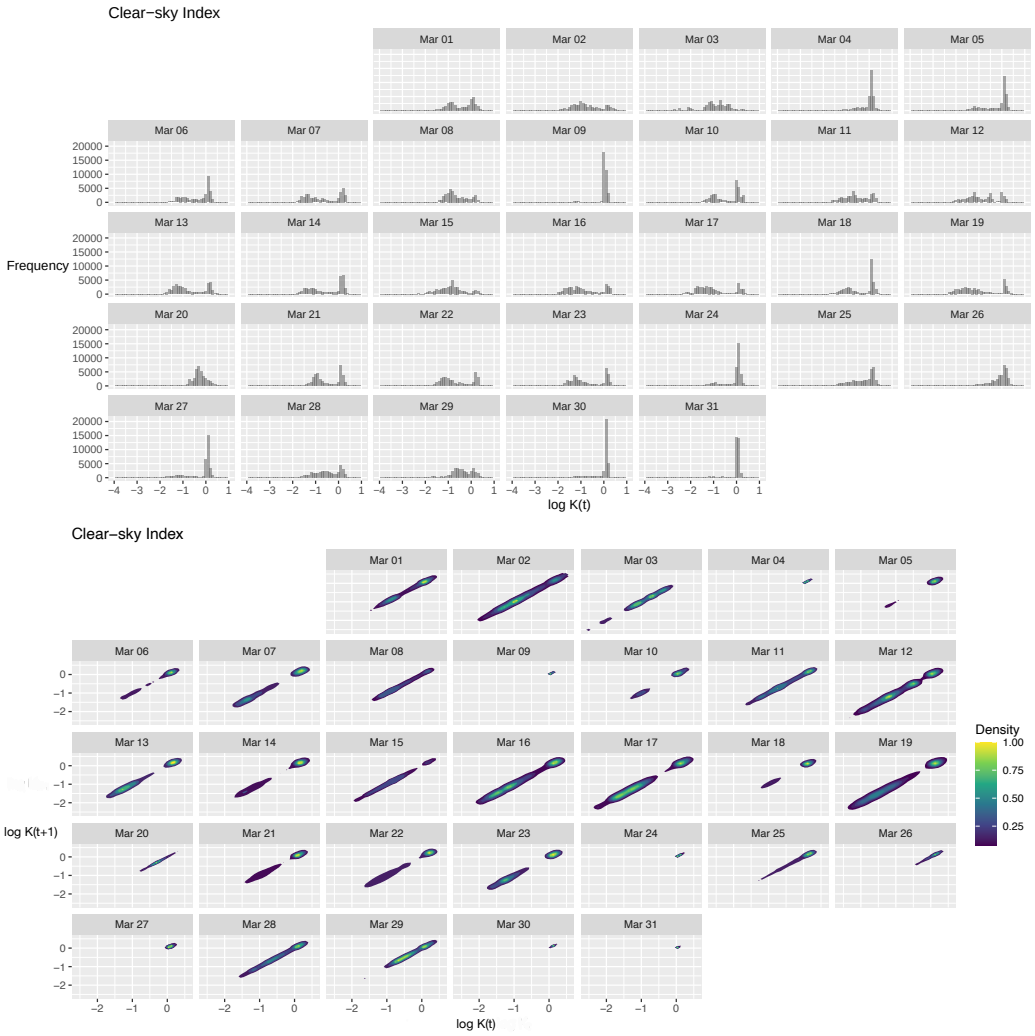


Figure 5: Empirical evidence against linearity. Top: Daily histogram of logarithm of clear-sky index time series. A bimodal histogram is a strong evidence against Gaussianity. Bottom: Bivariate density contour plots of $(\log K(t), \log K(t+1))$. The presence of two modes with different correlation structure supports the motivation on a two-state time series model.

model is reasonable. Two of the most commonly used are the nonlinearity test proposed by Keenan (1985) and Tsay (1986). Keenan's test is motivated by the approximation of a nonlinear stationary time series using a Volterra expansion. Assume the following model,

$$X(t) = \theta_0 + \sum_{i=1}^{p} \phi_i X(t-i) + \exp\left\{ \eta \left(\sum_{j=1}^{p} \phi_j X(t-j) \right)^2 \right\} + \varepsilon(t),$$

where ε_t are iid (0,1). Keenan's test is equivalent to the F-test for $\eta = 0$. It is computationally simple and handy on small sample sizes. Tsay's test extends Keenan's test with a more general nonlinear alternative by replacing the term $\exp\left\{ \eta \left(\sum_{j=1}^{p} \phi_j X(t-j) \right)^2 \right\}$ by

$$\exp\{\eta(\delta_{1,1} X(t-1)^2 + \delta_{1,2} X(t-1)X(t-2) + \ldots + \delta_{1,p} X(t-1)X(t-p) +$$
$$\delta_{2,2} X(t-2)^2 + \ldots + \delta_{2,p} X(t-2)X(t-p) + \delta_{m,m} X(t-m)^2)\}.$$

Both tests are implemented in the *TSA R package* (Chan and Ripley, 2018). We apply both tests to each of the daily log clear-sky index time series from Jan 21, 2011, to Dec 31, 2011, and find strong evidence against linearity (p-value < 0.01).

2.3 TAR model

In general, linear models have been shown to be very useful to predict weakly stationary processes. However, under the evidence of non-normality or nonlinearity, they can perform poorly. In nonlinear time series models, one possibility is an autoregressive polynomial model of higher order degrees. However, such models are not useful for extrapolating, since they quickly blowup to infinity. In this paper, we propose to use the threshold autoregressive (TAR) model to describe the dynamics of irradiance time series. We briefly introduce here the TAR model; more details can be found in Cryer and Chan (2008) and Tong (2010).

Tong and Lim (1980) proposed the TAR model as a piecewise linearization through the introduction of a switching mechanism. A general form of the TAR model assumes that

$$X(t) = a_{0,J_t} + \sum_{i=1}^{p} a_{i,J_t} X(t-i) + b_{J_t}\varepsilon(t),$$

where ε_t are iid (0,1) and $\{J_t\}$ is an indicator time series that takes values between $\{1, 2, \ldots, J\}$. To simplify this form, the most commonly used model is the two regime self-exciting threshold autoregressive (SETAR) model where $J_t = 1$ if $X(t-d) \leq \tau$ and $J_t = 2$ if $X(t-d) > \tau$. With this model, the parameter τ is the threshold parameter, and d is the delay parameter. Then, we refer to the SETAR model of order p to be

$$X(t) = \begin{cases} \phi_{1,0} + \sum_{j=1}^{p} \phi_{1,j} X(t-j) + \sigma_1 \varepsilon(t) & \text{if } X(t-d) \leq \tau; \\ \phi_{2,0} + \sum_{j=1}^{p} \phi_{2,j} X(t-j) + \sigma_2 \varepsilon(t) & \text{if } X(t-d) > \tau, \end{cases} \tag{2}$$

where $\{\phi_{i,j}, i = 1, 2,\ j = 0, 1, \ldots, p\}$ are real constants and ε_t are iid (0,1). When dealing with time series models, we need to question the existence of the stationary distributions for the process. Chan and Tong (1985) showed the sufficient conditions for a SETAR model to be ergodic and therefore the existence of a stationary distribution. However, these Lyapunov conditions are difficult to verify. In practice, a handy tool is to run the skeleton model (noise-free model) with different initial conditions (Tong, 2010).

Amendola et al. (2009) showed that the SETAR model with two regimes (2) is weakly stationary if the matrices $\Phi^{(1)}$ and $\Phi^{(2)}$ have both dominant eigenvalues less than one, where

$$\Phi^{(1)} = \begin{pmatrix} \phi_{1,1} & \cdots & \phi_{1,p} \\ \mathbb{I}_{p-1} & & \mathbb{O}_{[p-1]\times 1} \end{pmatrix} \quad \text{and} \quad \Phi^{(2)} = \begin{pmatrix} \phi_{2,1} & \cdots & \phi_{2,p} \\ \mathbb{I}_{p-1} & & \mathbb{O}_{[p-1]\times 1} \end{pmatrix},$$

$\mathbb{I}_{p-1}$ and $\mathbb{O}_{[p-1]\times 1}$ denote the identity matrix of dimension $p-1$ and a matrix of zeros of dimension $p-1\times 1$, respectively.

SETAR models have been widely used in financial time series analysis (Tong, 2010). The threshold model has also been applied to study epidemiological time series (Watier and Richardson, 1995) and cyclical fish landings (Samanta et al., 2011). These studies showed the versatility of SETAR nonlinear time-series models, which is capable of better describing cyclical fluctuations when there is evidence of non-linearity.

3 Probabilistic Modeling of the Clear-sky Index

We propose a new approach to model the irradiance time series. Our model does not require a priori classification of days among cloudy or bright days, which makes the statistical analysis more robust.

3.1 Model

Let $\log K_r(t)$ be the logarithm of clear-sky time series defined in (1) for day r at time t, where $r=1,\ldots,n$ and $t=1,\ldots,T$. We propose the multi-day SETAR model as follows:

$$\log K_r(t)=\begin{cases}\phi_{1,0}+\sum_{j=1}^{p}\phi_{1,j}\log K_r(t-j)+\sigma_{1,r}\varepsilon_r(t) & \text{if } \log K_r(t-d)\leq\tau;\\ \phi_{2,0}+\sum_{j=1}^{p}\phi_{2,j}\log K_r(t-j)+\sigma_{2,r}\varepsilon_r(t) & \text{if } \log K_r(t-d)>\tau,\end{cases} \tag{3}$$

where $\{\phi_{i,j}, i=1,2,\ j=0,1,\ldots,p\}$ are real constants, $\{\sigma_{i,r}, i=1,2, r=1,\ldots,n\}$ are positive constants, $\varepsilon_r(t)$ are iid (0,1), τ is a real value and d is a positive integer number.

Remark 1. Usually, time series analysis is performed with one single long (preferable) time series. However, in some applications, the available data consist of replicated series

(Silva et al., 2005). Our model assumes that each day is a replicate of a similar SETAR model. The model has the same autoregressive coefficients each day, although they may have different noise variations, $\{\sigma_{i,r}, i = 1, 2\}$.

Since irradiance data can only be recorded during the day time with sunlight (e.g., from 7 am to 5 pm), it is then not reasonable to join the daily time series as a uniquely long time series. Instead, our model assumes that, for each day, the observed time series is a replicate of a similar nonlinear cyclical time series, which we model as a SETAR model. To allow for different stochastic variations on irradiance data across days, which occurs due to different environmental conditions, we allow $(\sigma_{1,r}, \sigma_{2,r})$ to be day-specific.

Remark 2. Note that the multi-day SETAR model has the same skeleton model (noise free) for all days. Then, the conditions shown by Amendola et al. (2009) for the SETAR model to be stationary are also sufficient conditions for the multi-day SETAR model to be stationary per day.

3.2 Estimation procedure

Although we know the conditions for the stationarity of the multi-day SETAR model that can be used to compute the full likelihood, the stationary distribution does not have a closed form as in the SETAR model and it involves computing multiple integrals. Similar to the estimation of a SETAR model, we propose a two-step inference procedure. First, we estimate the threshold and delay parameters, τ and d, by using log-likelihood criteria. We then consider a weighted least squares method to estimate the autoregressive parameters, $\{\phi_{i,j}, i = 1, 2,\ j = 0, 1, \ldots, p\}$.

Estimation of d, τ and p

We temporarily assume that p is known and assume normality on all $\varepsilon_r(t)$. Then, for

each day r and value (d, τ), let $M_{1,r} = \{t : \log K_r(t-d) \leq \tau\}$ and $\log K_{1,r}(t)$ be the irradiance values $\log K_r(t)$ with $t \in M_{1,r}$. Similarly, we define $M_{2,r} = \{t : \log K_r(t-d) > \tau\}$ and $\log K_{2,r}(t)$ be the irradiance values $\log K_r(t)$ with $t \in M_{2,r}$. Denote by $m_{i,r}$ the cardinality of $M_{i,r}$, and $m_{1,r} + m_{2,r} = T - p$ for all r. For day r, we regress the series of $\log K_{1,r}(t)$ on its lags 1 to p to find estimates of $\hat{\phi}^r_{1,0}, \ldots, \hat{\phi}^r_{1,p}$ and compute the maximum likelihood noise variance estimate, $\hat{\sigma}^2_{1,r}$, by the sum of squared residuals divided by $m_{1,r}$. Similarly, we estimate $\hat{\sigma}^2_{2,r}$ using the series of $\log K_{2,r}(t)$. Then, we estimate (d, τ) by maximizing the profile conditional log-likelihood function

$$l(d,\tau) = -\frac{(T-p)n}{2}\{1 + \log(2\pi)\} - \sum_{r=1}^{n} \frac{m_{1,r}}{2} \log(\hat{\sigma}^2_{1,r}) - \sum_{r=1}^{n} \frac{m_{2,r}}{2} \log(\hat{\sigma}^2_{2,r}). \tag{4}$$

In practice, the autoregressive order p can be different for the two regimes. If this is the case, let p_1 and p_2 be the autoregressive orders for the low and high regimes, respectively, and estimate the parameters (d, τ, p_1, p_2) by minimizing the AIC, where

$$AIC(d, \tau, p_1, p_2) = -2l(d, \tau) + 2(p_1 + p_2 + 2).$$

Remark 3. For each day r, $\{\hat{\phi}^r_{1,0}, \ldots, \hat{\phi}^r_{1,p}, \hat{\phi}^r_{2,0}, \ldots, \hat{\phi}^r_{2,p}\}$ are consistent estimators for the parameters $\{\phi_{i,j}, i = 1, 2,\ j = 0, 1, \ldots, p\}$. However, a pooled estimator for all days $r = 1, \ldots, n$ will have lower variance, i.e, it will be more efficient.

Estimation of $\{\phi_{i,j}, i = 1, 2,\ j = 0, 1, \ldots, p\}$

To estimate the autoregressive coefficients we consider a conditional least squares (CLS) approach using the information across all days. The CLS estimator is obtained by mini-

mizing

$$Q(\boldsymbol{\phi}) = \sum_{r=1}^{n} \sum_{t \in M_{1,r}} \left(\log K_r(t) - \phi_{1,0} - \sum_{j=1}^{p} \phi_{1,j} \log K_r(t-j) \right)^2 + \sum_{r=1}^{n} \sum_{t \in M_{2,r}} \left(\log K_r(t) - \phi_{2,0} - \sum_{j=1}^{p} \phi_{2,j} \log K_r(t-j) \right)^2 ,$$

where $\boldsymbol{\phi}$ denotes the vector or autoregressive coefficients for both regimes. Note that we assume different variances $\sigma^2_{1,r}$ and $\sigma^2_{2,r}$ within each day r. The CLS estimator will then be unbiased but not necessarily the one with the least variance, and therefore not optimal. Our setting is very similar to the analysis of linear models with m different groups where variances are constant within each group. Then, weighted least squares (WLS) is the natural estimation method to apply. Hooper (1993) studied the problem of optimal weighted least squares for the estimation procedure.

If $\sigma^2_{1,r}$ and $\sigma^2_{2,r}$ are known for each r, a feasible solution is to apply WLS using weights $1/\sigma^2_{i,r}$, since these weights correspond to the optimal solution when all noise terms are normally distributed. However, we do not know the true variability within each day. Empirically, the effect of only using the MSE as estimators for the weights produces a poor estimation of the AR coefficients by shrinking too much of the data. This issue was already observed on Hooper (1993) paper and we also observe the poor estimation in our model and propose a similar strategy. Hooper (1993) shows that a WLS estimator with a set of weights that combines a model-free and model-based estimators for the variances is asymptotically optimal (unbiased and lowest variance). We adjust this technique to our model setting as follows.

The idea of weighted CLS (WCLS) estimation is to transform the variable of interest by multiplying each term with a weight $w_{i,r}$ and then apply the CLS to estimate the parameters

$\boldsymbol{\phi}$. Our proposed weights combine two elements: 1) the maximum likelihood noise variance estimate for day r, $\hat{\sigma}^2_{1,r}$ and $\hat{\sigma}^2_{2,r}$, and 2) the Bayesian estimate of noise variance using a normal model (O'Hagan and Forster, 2004). Our model-based estimator for the variance is the Bayesian estimator formulated as follows. For each i and r, the prior distribution for $1/\sigma^2_{i,r}$ is Gamma(a, b) and the prior distribution for $(\phi^r_{i,0}, \ldots, \phi^r_{i,p})'|\sigma^2_{i,r}$ is N$(0, \sigma^2_{i,r} I_{p+1})$, where $a > 0$, $b > 0$, I_{p+1} denotes the identity matrix of dimension $p + 1$ and $'$ denotes the transpose. Then, the corresponding posterior distribution is $\sigma^2_{i,r}|\mathbf{X}, \mathbf{Y} \sim IG(\frac{\gamma_{i,r}}{2}, \frac{2}{\gamma_{i,r}\theta_{i,r}})$, where $\mathbf{X}, \mathbf{Y}$ denotes all the data information, past and present (analogous to linear regression models), and IG denotes the inverse-gamma distribution. The posterior parameters are

$$\gamma_{i,r} = 2 * (a + m_{i,r}/2) \qquad \theta_{i,r} = \frac{1}{(a + m_{i,r}/2)(b + (\mathbf{Y}'\mathbf{Y} + \boldsymbol{\mu}'\Lambda\boldsymbol{\mu})/2)},$$

where $\Lambda = (\mathbf{X}'\mathbf{X} + I_{p+1})^{-1}$ and $\boldsymbol{\mu} = \Lambda\mathbf{X}'\mathbf{Y}$.

By considering the previous model and applying Theorem 3 in Hooper (1993), we obtain the optimal weights as

$$w_{i,r} = \frac{m_{i,r} + \gamma_{i,r}}{m_{i,r}\hat{\sigma}^2_{i,r} + \gamma_{i,r}\theta_{i,r}}.$$

Lastly, the WCLS estimator for $\boldsymbol{\phi}$ is the one that minimizes:

$$\begin{aligned}\tilde{Q}(\boldsymbol{\phi}) = &\sum_{r=1}^{n} \sum_{t \in M_{1,r}} \left(\log Z_r(t) - \phi_{1,0} w_{1,r} - \sum_{j=1}^{p} \phi_{1,j} \log Z_r(t-j) \right)^2 \\ &+ \sum_{r=1}^{n} \sum_{t \in M_{2,r}} \left(\log Z_r(t) - \phi_{2,0} w_{2,r} - \sum_{j=1}^{p} \phi_{2,j} \log Z_r(t-j) \right)^2, \end{aligned} \tag{5}$$

where $Z_r(t) = w_{1,r} \log K_r(t)$ if $t \in M_{1,r}$ and $Z_r(t) = w_{2,r} \log K_r(t)$ if $t \in M_{2,r}$.

Compared to the one day estimator, the WCLS estimator $\hat{\boldsymbol{\phi}}_{WCLS} = \text{argmin}_{\boldsymbol{\phi}}\, \tilde{Q}(\boldsymbol{\phi})$ has

a reduced variance by a factor $1/n$ because of available replicates.

4 Statistical Analysis of Guadeloupe Island's Data

As mentioned in Section 2, we analyze the log clear-sky index, $\log K(t)$, from Jan 21, 2011, to Dec 31, 2011, corresponding to the Guadeloupe island group. The considered period for all days is from 7 : 00 hrs to 17 : 00 hrs, and data are collected every second. We have then a total of 345 daily time series, each consisting of 36,000 measurements. We divide the complete data set into a training set and a testing set. We fit our multi-day model (3) to the first 300 days to describe the stochastic variability of the irradiance time series. Then, we compare the prediction using our fitted model for the last 45 days.

4.1 Estimation of (d, τ, p_1, p_2)

First, we choose the threshold value τ, delay parameter d, and autoregressive orders p_1 and p_2 as described in the previous section. Figure 6 shows the AIC values for different combinations of the parameter space. We search for the parameter τ within the 0.15% and 0.85% quantile of the observed values. The smallest value for d gives lower values for the AIC function and a threshold value within $(-.1, 0)$ for all different combinations of the autoregressive orders. By increasing the autoregressive order of the high regime, p_2, we also decrease the AIC values. In contrast, there is no significant decrement when increasing the autoregressive order of the low regime, p_1. Then, we get $\hat{p}_1 = 4$, $\hat{p}_2 = 5$, and by minimizing the AIC we choose $\hat{d} = 1$ and $\hat{\tau} = -0.02$.

If $\log K(t) = 0$, then the observed irradiance is equal to the expectation that corresponds to a clear sky period. Then, the estimated threshold value $\hat{\tau} = -0.02$ can be interpreted as a split between a clear sky regime (high regime, $\log K(t) > -0.02$) and non-clear sky regime (low regime, $\log K(t) \leq -0.02$) due to environmental conditions.

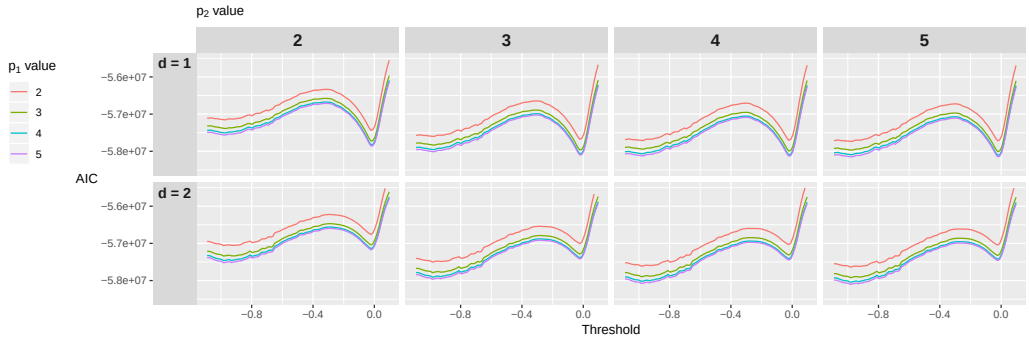


Figure 6: Estimation of (d, τ, p_1, p_2) via minimum AIC. To reach the minimum AIC the smaller value of $d = 1$ is preferable.

4.2 Estimation of noise variability and autoregressive parameters

We fit a SETAR model per each day r with $(\hat{d}, \hat{\tau}, \hat{p}_1, \hat{p}_2)$ and estimate the across days variabilities $\sigma^2_{i,r}$ $i = 1, 2$ for the non-clear sky regime and clear sky regime, respectively. Now, we use these values together with the model-based estimators to compute the weighted least squares estimator of the autoregressive parameters $\{\phi_{j,i},\ j = 1, 2, 3, 6,\ i = 1, 2\}$. For the Bayesian normal-mixture model, we fix the hyper parameters $a = 0.01$ and $b = 0.01$. We update the estimated variances $\hat{\sigma}^2_{i,r}$ values by computing the squared residuals with the WCLS estimators. Figure 7 shows the distribution of the estimated variability across days, $\sigma^2_{1,r}$ and $\sigma^2_{2,r}$. The noise variability can be higher in the non-clear sky regime than in the

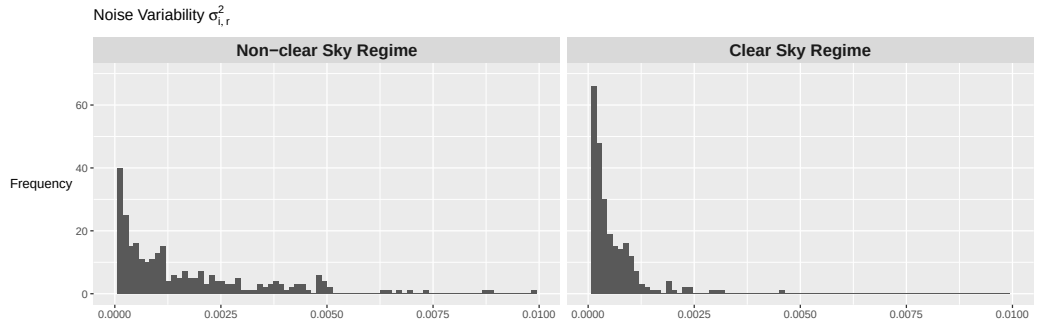


Figure 7: Estimated variability across days for non-clear sky regime ($\log K_r(t-1) \leq -0.02$) and clear sky regime ($\log K_r(t-1) > -0.02$), respectively. Noise variability is higher under non-clear sky regime.

clear-sky regime. Lastly, our estimated multi-day model is:

$$\begin{aligned}\log K_r(t) =&(0.00024 + 1.538 \log K_r(t-1) - 0.614 \log K_r(t-2) + 0.120 \log K_r(t-3) \\ &- 0.045 \log K_r(t-4) + \sigma_{1,r}\varepsilon_r(t))\mathbb{1}\{\log K_r(t-1) \leq -0.02\}+ \\ &(-0.00051 + 1.601 \log K_r(t-1) - 0.790 \log K_r(t-2) + 0.240 \log K_r(t-3) \\ &- 0.092 K_r(t-4) + 0.040 \log K_r(t-5) + \sigma_{2,r}\varepsilon_r(t))\mathbb{1}\{\log K_r(t-1) > -0.02\},\end{aligned} \tag{6}$$

where $\{\sigma_{i,r}, i = 1, 2, r = 1, \ldots, n\}$ are positive constants, $\varepsilon_r(t)$ are iid (0,1) and $\mathbb{1}\{A\}$ is equal to 1 if A is true.

4.3 Uncertainty of parameter estimates

Based on our 2-step estimation procedure, quantifying the uncertainty propagation is a challenge. We investigate if the small changes in the selection of the threshold/delay parameter could produce significant changes on the autoregressive parameter estimates. We

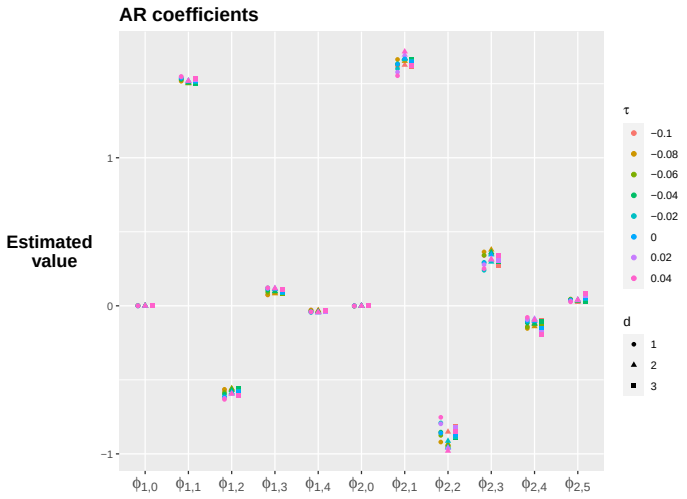


Figure 8: Results of numeric study to explore the effect of τ and d selection to the autoregressive parameters estimates $\{\phi_{i,j}\}$. Each point corresponds to an estimated autoregressive parameter given τ and d (p_1 and p_2 are fixed).

perform a small numeric experiment, where we select threshold values, τ, between $-.01$ and .04, and $d = 1, 2, 3$. Then, we repeat the estimation of the autorregresive parameters as in section 4.2. Figure 8 shows that there is not a strong influence. We observe variability on the estimated parameters but the effect is very small in most cases. It might be because we have a large data set, so the autoregressive estimated parameters are consistent.

Now, we assume our estimated model is the ground truth and generate bootstrap series as follows. 1) Sample a paired value of $(\sigma_1, r, \sigma_{2,r})$. 2) Start $K(t) = 0$ for $t = 1, \ldots, 5$. 3) Using Gaussian noise, we generate a series of lengths $T = 36000$ following model 6. 4) Repeat steps 1-3, 300 times. The results of 1-4 are bootstrapped data series of the same length and days as the original data. 5) Use the bootstrapped data set to estimate $\hat{\phi}^B_{i,j}$. We repeat this procedure $M = 500$ times. Figure 9 shows the boxplots of the estimated parameters. Overall, the uncertainty is low, being the non-clear sky regime of higher

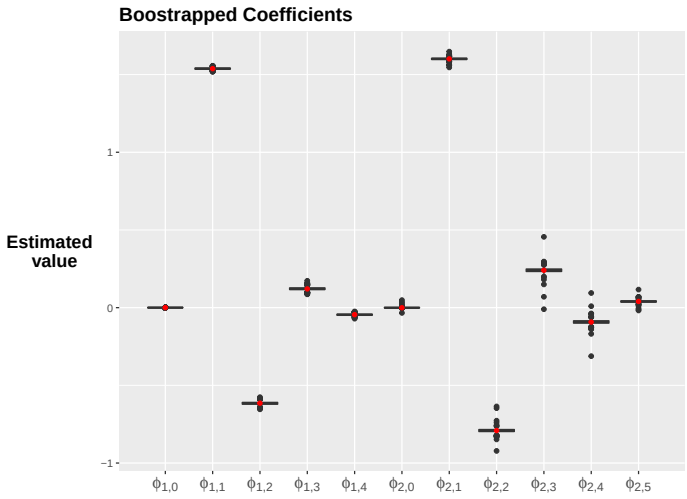


Figure 9: Bootstrap results for parameters estimates $\{\phi_{i,j}\}$. Each point corresponds to an estimated autoregressive parameter using a bootstrapped sample and given $\tau = -0.02$, $d = 1$, $p_1 = 4$ and $p_2 = 5$.

estimation uncertainty compare to the clear-sky regime.

4.4 Prediction accuracy of the estimated multi-day SETAR model

We use our estimated model (6) to obtain one step forecasted values. Assume for day r that we have information up to time t. Then, the one-step-ahead forecasted value is $\log \widehat{K_r(t+1)} = 0.00024 + 1.538 \log K_r(t) - 0.614 \log K_r(t-1) + 0.120 \log K_r(t-2) - 0.045 \log K_r(t-3)$ if the latest value of $\log K_r(t)$ belongs to the non-clear sky regime (≤ -0.02) or $\log \widehat{K_r(t+1)} = -0.00051 + 1.601 \log K_r(t) - 0.790 \log K_r(t-1) + 0.240 \log K_r(t-2) - 0.092 K_r(t-3) + 0.040 \log K_r(t-4)$ in the other case. For $h > 1$, the h-step-ahead prediction follows the same rule but replaces the $\log K_r(t+h-1)$ by its predicted value $\log \widehat{K_r(t+h-1)}$.

We apply this rule to the training and testing data sets. Figure 10 shows the histogram

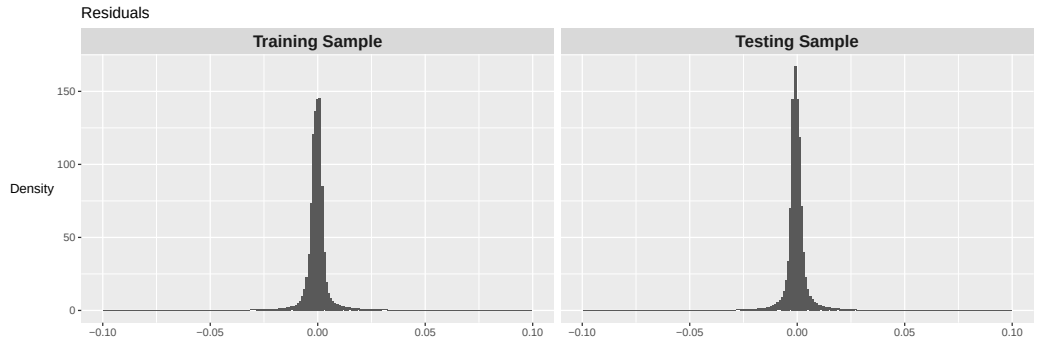


Figure 10: Residuals computed using the one-step ahead forecasted and observed values. Residuals are very small in both the training and testing data sets.

of the residuals computed between the observed and forecasted values, $\log K_r(t+1) - \log \widehat{K_r(t+1)}$. In both samples, training and testing the residuals are close to zero, suggesting a highly accurate fitting of the model to the real data. Figure 11 shows a six-days subsample (three on the training and three on the testing sets) of the observed and predicted series. The fitting of our model is highly accurate for both clear sky and non-clear sky periods.

Additionally, we explore the uncertainty in prediction using the bootstrapped samples obtained in Section 4.3. We compute 10-sec-ahead predictions on some of the testing days using the bootstrapped values $\hat{\phi}^B_{i,j}$. Then, we built prediction bands using the .025 and .0975 empirical quantiles. Figure 12 shows segments of randomly selected irradiance series to illustrate better the uncertainty bands. Observed data is indicated using black lines and dots, blue dots indicate predictions, and the uncertainty prediction bands are red. The forecast is close to the actual values, even though the observed data is sometimes outside the uncertainty bands.

An interesting question is whether our method is more accurate than other prediction focus methods for solar irradiance data. We compare our results with those predictions

Figure 11: Day examples of model prediction using the one-step-ahead forecast rule. First row corresponds to dates on the training set and second row corresponds to dates on the testing set. Observed and forecasted data are shown in black and red, respectively. The prediction performance is very accurate in both cases (for training and testing dates).

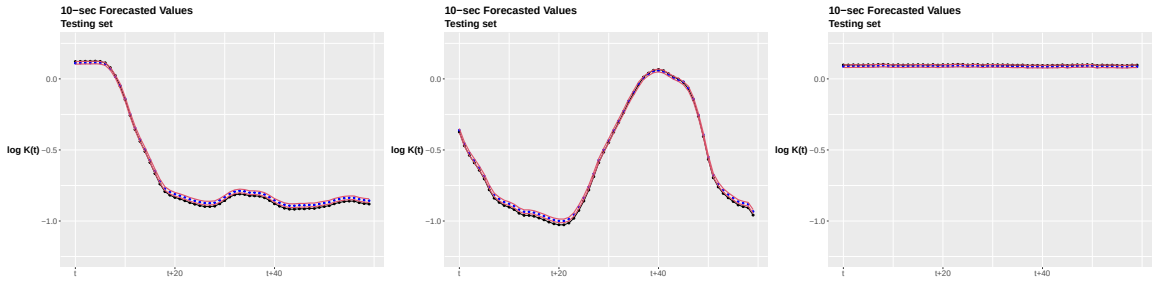


Figure 12: 10-step-ahead predictions illustration. Blue points are the 10-sec-ahead predictions and the black line with points are the observed data. Prediction bands (red) are computed using 500 bootstrap samples.

obtained by the PAR model, ARIMA model, and two machine learning algorithms, the multilayer perceptron (MLP) and the Bagged Regression Trees (RT). Fouilloy et al. (2018) shows a comparative study of machine learning forecast methods under different weather variability factors. Among these methods, the MLP and Bagged RT were the most accurate methods. Below, we describe briefly the competitors methods applied to $\log K_r(t)$.

1. **PAR model** assumes that $\log K_r(t) = \sum_{i=1}^{p_t} \phi_i(t) \log K_r(t-i) + \varepsilon_r(t)$, where $\varepsilon_r(t)$ is a periodic white noise with variance $\sigma^2(t)$. In other words, the PAR model assumes an autoregressive model where the coefficients depend on the day-hour, t. To fit the PAR model, we use least squares for each t using different days' information. Finally, the one-step-ahead forecast is

$$\widehat{\log K_r(t+1)} = \sum_{i=1}^{p_t} \phi_i(t) \log K_r(t+1-i).$$

2. **ARIMA(p,d,q) model** is the most general time series model which assumes that $\phi(B)(1-B)^d \log K(t) = \theta(B)\varepsilon(t)$, where B is the backshift operator and $\phi(z)$ and $\theta(z)$ are polynomials of degrees p and q respectively. Using AIC criteria we fit to the data an ARIMA(1,1,2) model to the irradiance series. More details and the h-step-ahead predictor see Brockwell and Davis (2006).

3. **Bagged RT** is an improvement on RT, the model creates B bootstrapped samples from the training set and finds the best (simple) regression tree within each sample. Then, we average all the predictions. This procedure reduces prediction variance. We use 25 added regression trees.

4. **MLP** is a type of Artificial Neural Networks with one hidden layer and one output layer used. The prediction of the MLP using m neurons, one output neuron and t

input variables is

$$\log \widehat{K_r(t+1)} = \sum_{j=1}^{m} w_j g \left(\sum_{i=0}^{t-1} \omega_{i,j} \log K(t-i) + b_j \right),$$

where $\omega_{i,j}$ are the weights between the input i and the neuron m, g is the activation function, and w_j is the weight between the output and the hidden neuron. Similar to Fouilloy et al. (2018), we select $m = 5$ hidden neurons and the sigmoid activation function.

For h-step-ahead forecast where $h > 1$, we apply the same rule as with the SETAR model. We compare the 5 methods using the Mean Squared Error (MSE),

$$MSE(h) = \sum_{r} \sum_{t} \left(\log K_r(t+h) - \log \widehat{K_r(t+h)} \right)^2.$$

We use the first 300 days to train the models and compare the forecasts values on the last 45 days. To fit the ARIMA, Bagged RT and MLP we use the *R* packages *forecast, ipred* and *neuralnet* (Hyndman and Khandakar, 2008; Peters and Hothorn, 2019; Fritsch et al., 2019).

Table 1 shows the MSE values for 1-sec-ahead and 5-sec-ahead prediction forecast. We include the standard deviation of the SE (squared error) in brackets. In both cases, the multi-day SETAR model has the smallest MSE. For the 1-sec prediction, the closest competitor is the PAR model while for the 5-sec prediction the closest competitors are the ARIMA and MLP. Notice that PAR 5-sec predictions are considerably high compared to the other competitors. Based on data results, the PAR coefficients are not robust to large drops on the irradiance series and highly affect the long-term prediction. Overall, the results agree with what we expected due to the highly non-linearity observed on the data. We conclude that our model is preferable since the prediction are as good as other competitors and our

model has a stronger interpretability in terms of clear-sky and non-clear sky regimes.

Method	1-step-ahead forecast	5-step-ahead forecast
Multi-day SETAR	**0.00037** (0.0022)	**0.01409** (0.1006)
PAR model	0.00048 (0.0006)	206.97 (401.90)
ARIMA(1,1,2)	0.00616 (0.06177)	0.01421 (0.0970)
Bagged RT	0.02455 (0.0680)	0.03490 (0.1096)
MLP	0.00615 (0.0609)	0.01546 (0.0970)

Table 1: MSE Error in the h-step-ahead forecast with standard deviation of SE in brackets. Smallest values are highlighted in bold.

4.5 Stochastic features of the estimated multi-day SETAR model

In this section, we investigate some probabilistic features of our fitted model. Our goal is to identify characteristics of the clear-sky and non-clear sky regimes that could be used for the grid manager to determine in which regime the grid is working. First, we verify if the estimated model corresponds to a stationary model for each day. To do this, we compute the eigenvalues of the matrices

$$\Phi_1 = \begin{pmatrix} 1.538 & -0.614 & 0.120 & -0.045 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \text{ and } \Phi_2 = \begin{pmatrix} 1.601 & -0.790 & 0.240 & -0.092 & 0.040 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

Then, the maximum eigenvalue for Φ_1 is 0.9975 and the maximum eigenvalue for Φ_2 is 0.9969. Following Amendola et al.'s criteria, the estimated model corresponds to a stationary model. We also consider the skeleton plot as suggested by Tong (2010). The skeleton

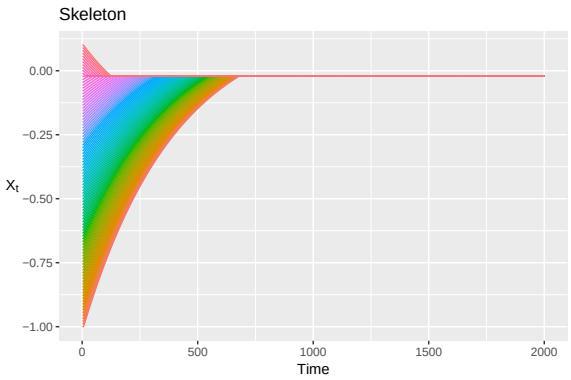


Figure 13: Skeleton of the estimated multi-day SETAR model with different initial values. The skeleton of the estimated model is bounded when T increases for different initial values which is a graphical tool of ergodicity.

corresponds to the deterministic part of the model, i.e.,

$$\begin{aligned}\log K_r(t) = (&0.00024 + 1.538 \log K_r(t-1) - 0.614 \log K_r(t-2) + 0.120 \log K_r(t-3) \\ &- 0.045 \log K_r(t-4)) \mathbb{1}\{\log K_r(t-1) \leq -0.02\} + \\ &(-0.00051 + 1.601 \log K_r(t-1) - 0.790 \log K_r(t-2) + 0.240 \log K_r(t-3) \\ &- 0.092 K_r(t-4) + 0.040 \log K_r(t-5)) \mathbb{1}\{\log K_r(t-1) > -0.02\}.\end{aligned}$$

We start the skeleton with different initial values and follow the trajectory. If the skeleton is bounded when t increases, then it is visual proof of ergodicity of the process. Figure 13 plots some paths of the skeleton, and it also suggests that our model is stationary. Then, we have enough evidence to claim that our estimated multi-day SETAR model is stable. Therefore, we can compute a stationary form of the autocovariance and spectral density

Now, we can compare the characteristics of the clear sky and non-clear sky regimes. Figure 14 (left) plots the autocorrelation function (ACF) computed with the coefficients for each regime. Here, a lag corresponds to one second. While the non-clear sky regime is

more dependent on its past values, the clear sky regime (if we stay on it) values are almost independent after 10 minutes (600 seconds). Additionally, we visualize the spectral density for each regime computed as

$$S_i(w) = \frac{\sigma^2_{i,r}}{|\phi_i(\exp(-\mathbf{i}2\pi w))|^2},$$

where $\phi_i(z) = 1 - \sum_{j=1}^{p} \phi_{i,j} z^j$ and $\mathbf{i} = \sqrt{-1}$. The spectrum gives information about the oscillatory behavior of the series. If the spectrum is concentrated in the low frequencies, it represents a slower oscillation with a longer period and an amplitude larger than that of a signal with a spectrum concentrated in high frequencies. Figure 14 (right) shows the spectrum for each regime computed using the mean variance per regime. The non-clear sky regime shows a higher power on the lowest frequencies in comparison with the clear sky regime, suggesting slower but stronger oscillations in the non-clear sky regime. This is in agreement with the fact that we expect higher drops on the irradiance values when we are on a period of the non-clear sky regime.

Remark 4. Note that the spectrum explains the variability of each regime, assuming a common constant variance for the noise term that represents the dominating oscillatory behavior. When the noise for each day has a different variability, the oscillatory behavior will change accordingly.

Remark 5. Regime models can also be estimated using the oscillatory properties. For example Hadj-Amar et al. (2021) fitted a HMM assuming that periodicity characterized each regime. We could use a combination between these ideas and Whittle likelihood to develop an alternative inference procedure.

As mentioned in the introduction, it is useful to generate synthetic irradiance time series to quantify the variability of the data. We simulate a one-hour time series from our estimated multi-day model (6) to verify if we can replicate some of the commonly observed

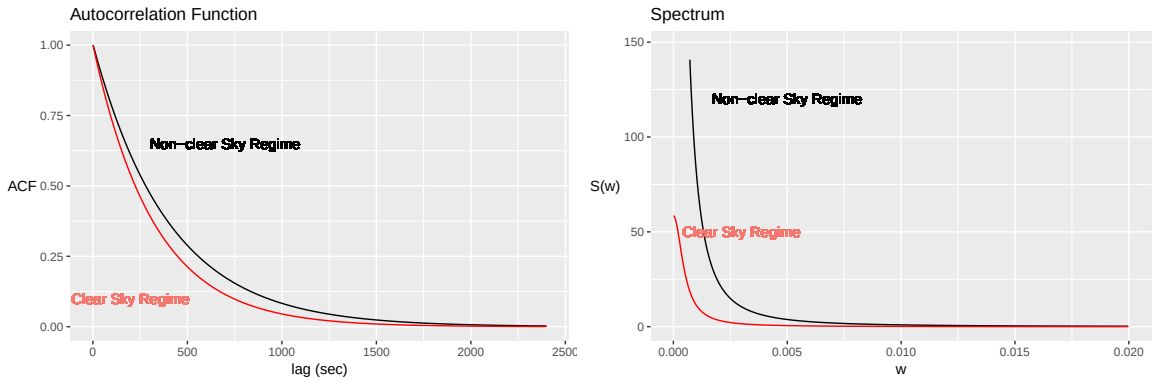


Figure 14: Comparison of the clear sky and non-clear sky regimes. The autocorrelation function suggest a stronger time dependence for the non-clear sky regime. The log-spectrum shows that the non-clear regime has a higher amplitude than the clear regime.

paths in the data. To simulate from model (6) we need a pair value of $(\sigma^2_{1,r}, \sigma^2_{2,r})$. We randomly select an estimated pair value plotted in Figure 7 and then we run the model. Figure 15 shows four different simulated scenarios; the right plot corresponds to the different paired noise variability options, and the red dot corresponds to the selection. The first scenario simulates an hour with a very noisy first half hour but a bright sky on the last half hour. The third scenario can also reproduce a clear-sky period, although the non-clear sky period is not as strong (smaller drops) as the one obtained with the first scenario. In general, our model can capture different scenarios depending on the noise variability of the day. Another interesting question is whether these scenarios are related to the seasons of the year. We consider the estimated values for the noise variability and split them into four groups: Spring, Summer, Autumn and Winter. We then plot the boxplot per group, as shown in Figure 16. This figure suggests that longer periods of clear sky can be observed during Spring since the noise variability is smaller for both regimes during this season. In contrast, Summer and Winter are periods with a higher variability, and as a consequence, higher drops on irradiance values might be observed. This is a reasonable hypothesis since

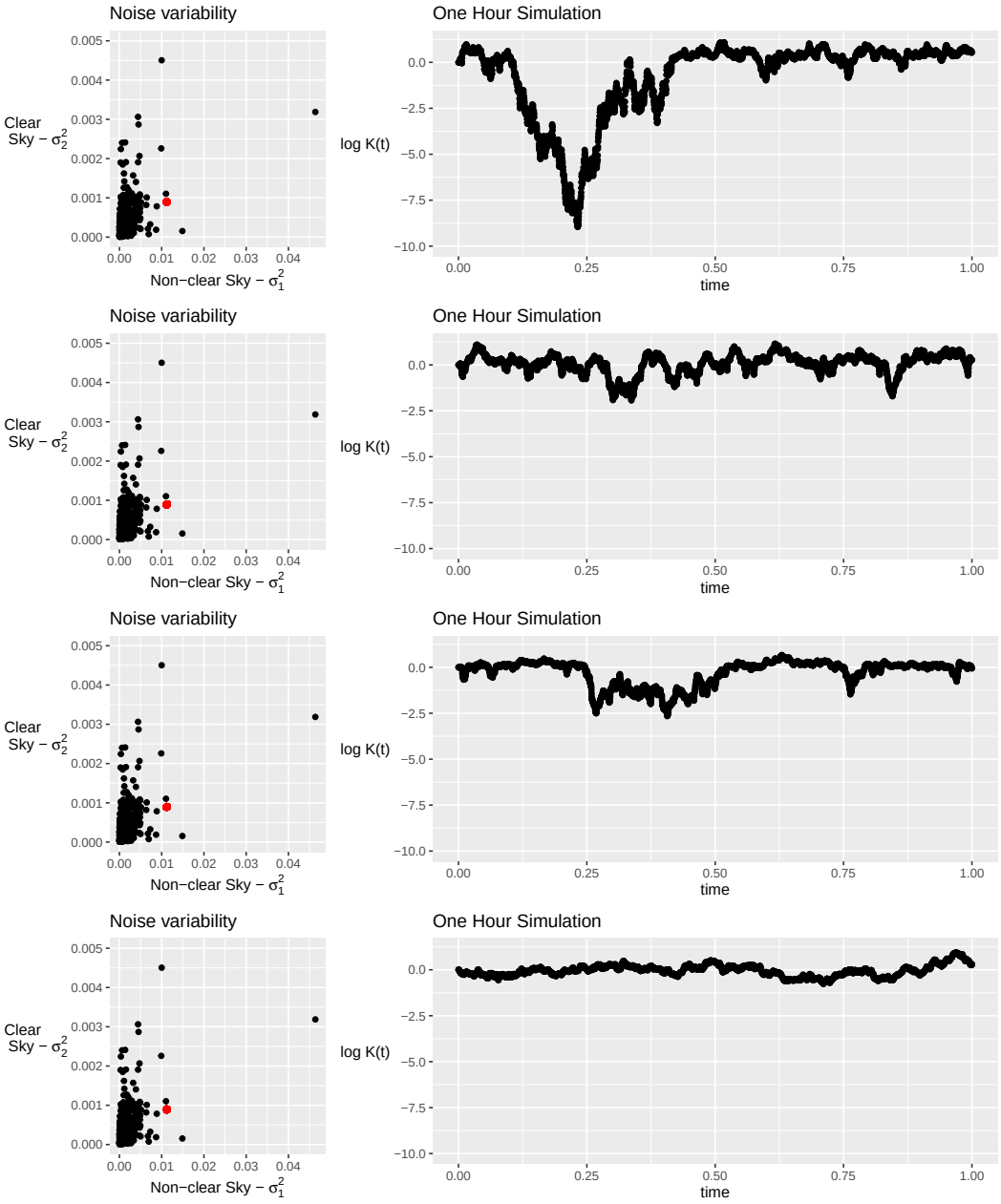


Figure 15: One-hour simulations of the multi-day model with different noise variabilities (shown in red). Depending on the noise variability, we observed long or short periods of clear sky regime.

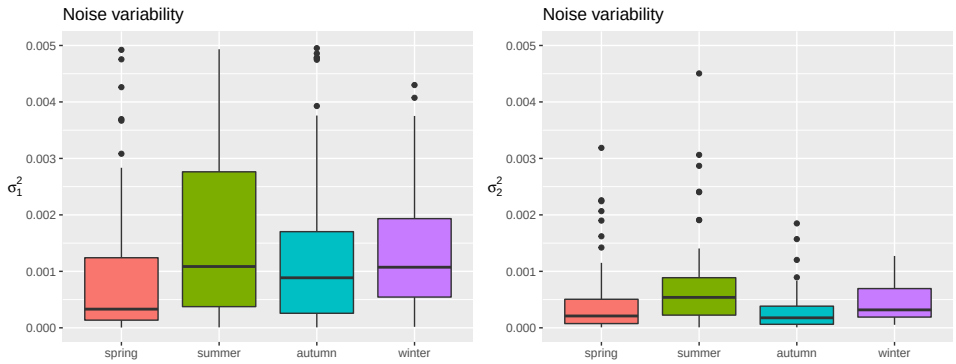


Figure 16: Boxplot of noise variability values per season: Spring, Summer, Autumn and Winter.

Summer and Winter are usually rainy seasons in the Caribbean sea.

4.6 Probabilistic forecasts using the multi-day SETAR model

Although one step prediction are very informative for the mean behave of the irradiance series, we can obtain more information about the irradiance variability by using probabilistic forecasts. We consider the following algorithm to generate a probabilistic forecast of log clear-sky index series. First, we sample a pair of value for the intraday variability $(\hat{\sigma}^2_{1,r}, \hat{\sigma}^2_{2,r})$. Then, given the series up to time t we simulate the future irradiance value as

$$\begin{aligned}\overline{\log K_r(t+1)} =&(0.00024 + 1.538 \log K_r(t) - 0.614 \log K_r(t-1) + 0.120 \log K_r(t-2) \\ &- 0.045 \log K_r(t-3) + \hat{\sigma}_{1,r}\varepsilon_r(t+1))\mathbb{1}\{\log K_r(t) \leq -0.02\}+ \\ &(-0.00051 + 1.601 \log K_r(t) - 0.790 \log K_r(t-1) + 0.240 \log K_r(t-2) \\ &- 0.092 K_r(t-3) + 0.040 \log K_r(t-5) + \hat{\sigma}_{2,r}\varepsilon_r(t+1))\mathbb{1}\{\log K_r(t) > -0.02\},\end{aligned}$$

where $\varepsilon_r(t+1) \sim N(0,1)$ for all r and t. For values of more than one step ahead, $\overline{\log K_{t+h,r}}$ $h > 1$, we use the previous simulated values $\overline{\log K_{t+h-1,r}}$ as if they are observed values and apply the same random generator.

To evaluate probabilistic forecasts, we use scoring rules. Scoring rules provide summary measures based on the predictive distribution and the event or observed value. Gneiting and Raftery (2007) defined the continuous ranked probability score (CRPS) as

$$CRPS(F, y) = \int (F(x) - \mathbb{1}\{x \geq y\}) \mathrm{d}x,$$

where F is the probability distribution corresponding to the probabilistic forecasts, and y is an observed value. In practice, we may not be able to know the forecasts probability distribution F. Then, we can use an estimator of the CRPS based on a generated sample and a true observed value. We compute the CRPS using the *scoringRules R package* (Jordan et al., 2017). There are no direct competitors to our model in the literature of application of time series models to irradiance data and perform probabilistic forecast. Then, we consider the Persistence Ensemble (PeEn) as the benchmark for the probabilistic forecast of irradiance data (Munkhammar et al., 2019). The idea of the PeEn method involves forecasting a future irradiance value based on the empirical distribution of the last s past values, i.e., we generate $\log K_r(t+1)$ by randomly select a value from the set $\{\log K_r(t), \ldots, \log K_{t-s,r}\}$.

In this experimental setting, we evaluate the performance for both methods, using the testing data set. For each day, we use the series until time $t = 12, 13, 14$ and 15 hours and we simulate 1000 forecasted 5-minute series $(t + 300)$. For the PeEn method, we use the 10-minute previous data. We then estimate the CRPS for each method and calculate the average per day. Figure 17 shows three different examples of irradiance data scenarios. The red dashed line corresponds to the PeEn method, and the continuous line corresponds to

our multi-day SETAR model. On a day where the irradiance series can be very variable, our model can characterize the variability better than its competitor. When abrupt changes do not occur, the multi-day SETAR model can describe the irradiance series as well as the PeEn method. The improved CRPS suggests that our model improves the density estimation of the irradiance time series. Finally, we conclude that our model can reproduce the random variability of irradiance series more accurately than its competitor.

5 Discussion

In this paper, we propose a multi-day SETAR model to forecast the irradiance series, and provide a clear estimation procedure for our model parameters that guarantees the consistency of the estimators. We also provide a detailed description of the interpretation of our model for irradiance series. Our model naturally classifies the two regimes as a clear-sky regime and non-clear sky regime, without any prior classification of the data. Then, we can study the stochastic features of each regime in detail. Our model can accurately forecast irradiance values, not only on the average but its variability as well. We did not face any computational difficulty to fit our model; however, the complexity is similar to fitting a linear model. It could be a trouble for larger data sets.

Accurate quantification of solar variability in a particular site is crucial to evaluate the solar-resource risk. These physical requirements will influence both technology development and financial performance. Our proposed model improves the variability quantification and reduces the bias of subjective classification for clear and non-clear days. Then we achieved a more robust and accurate quantification of resource-data uncertainty. Although we do not explore the online prediction, our model can achieve this purpose. The critical point is then the selection of noise variation. For a new day, the online prediction algorithm could start with a prior distribution of the noise variance. After one hour, an update of

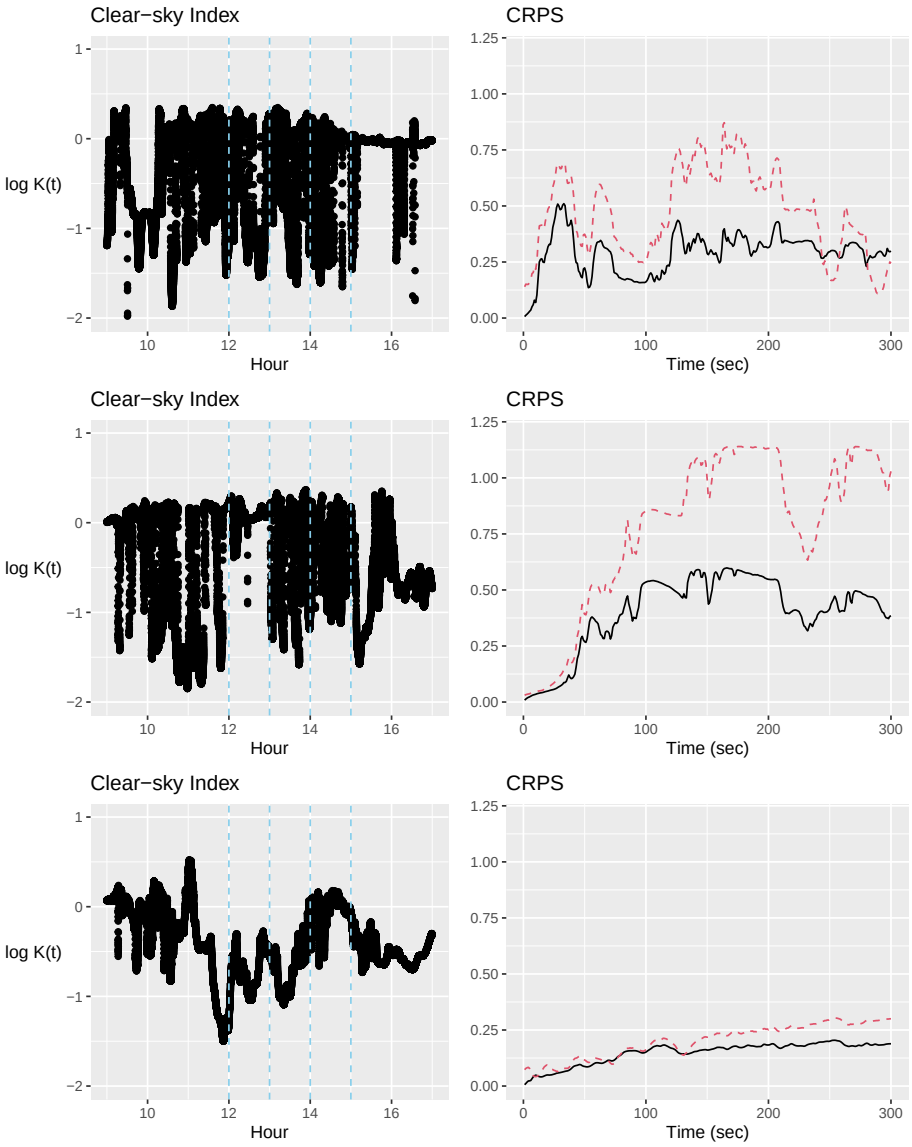


Figure 17: Computed CRPS. We compare our probabilistic forecast, which is based on the multi-day SETAR model (black continuous line), with the PeEn (red dashed line) method. Our method can forecast drops on irradiance series more accurately than the PeEn.

the noise variance can be done using the observed data and draw an update value for the noise variation needed by the probabilistic forecasts.

SETAR models are versatile models for non-linear time series. Our proposal, a multi-day SETAR model, uses the advantage of this family and the information provided across different days in a specific location. One unresolved challenge task is to forecast long periods of clear-sky state, but this can be improved if more information is available (e.g. external covariates or spatial neighbors). For example, covariates such as environmental conditions could be incorporate to model the threshold parameter τ or the noise variances. The estimation in such a case will be more challenging, and the proposed procedure here needs adaptations.

A future extension might be proposed using replicates across days and borrow information from different areas to forecast values in not observed nearby regions. A Spatial SETAR model could use nearby location values to propose a regime indicator variable. We should consider some specific features related to the data application: 1) Distances dependence; if we take two close locations, the irradiance is likely to be the same values, and two very far could be completely independent. Therefore a study of accurate location of sensors or selecting sites needs to be done. 2) Alternative, other spatial covariates might be more informative. The proposal of a regime indicator based on spatial covariates such as temperature or cloud movements might significantly impact prediction accuracy.

Funding

This publication is based upon work supported by King Abdullah University of Science and Technology (KAUST), Office of Sponsored Research (OSR) under Award No: OSR-2019-CRG7-3800

References

Al-Awadhi, S. A. and El-Nashar, N. (2002). Stochastic modelling of global solar radiation measured in the state of kuwait. *Environmetrics*, 13(7):751–758.

Amendola, A., Niglio, M., and Vitale, C. (2009). Statistical properties of threshold models. *Communications in Statistics - Theory and Methods*, 38(15):2479–2497.

Barbič, J., Safonova, A., Pan, J.-Y., Faloutsos, C., Hodgins, J. K., and Pollard, N. S. (2004). Segmenting motion capture data into distinct behaviors. In *Proceedings of Graphics Interface 2004*, page 185–194, Waterloo, CAN. Canadian Human-Computer Communications Society.

Boland, J., Scott, L., and Luther, M. (2001). Modelling the diffuse fraction of global solar radiation on a horizontal surface. *Environmetrics*, 12(2):103–116.

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

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

Brockwell, P. J. and Davis, R. A. (2006). *Time series: theory and methods*. Springer, New York. Reprint of the second (1991) edition.

Chan, K. and Tong, H. (1985). On the use of the deterministic lyapunov function for the ergodicity of stochastic difference equations. *Advances in Applied Probability*, 17(3):666–678.

Chan, K.-S. and Ripley, B. (2018). *TSA: Time Series Analysis*. R package version 1.2.

Cryer, J. D. and Chan, K.-S. (2008). *Time Series Analysis: With Applications in R*, chapter Threshold Models. Springer-Verlag New York.

Das, S., Genton, M. G., Alshehri, Y. M., and Stenchikov, G. L. (2021). A cyclostationary model for temporal forecasting and simulation of solar global horizontal irradiance. *Environmetrics*, page e2700.

Fernández-Peruchena, C. M., Blanco, M., Gastón, M., and Bernardos, A. (2015). Increasing the temporal resolution of direct normal solar irradiance series in different climatic zones. *Solar Energy*, 115:255 – 263.

Fouilloy, A., Voyant, C., Notton, G., Motte, F., Paoli, C., Nivet, M.-L., Guillot, E., and Duchaud, J.-L. (2018). Solar irradiation prediction with machine learning: Forecasting models selection method depending on weather variability. *Energy*, 165:620 – 629.

Fox, E. B., Hughes, M. C., Sudderth, E. B., and Jordan, M. I. (2014). Joint modeling of multiple time series via the beta process with application to motion capture segmentation. *The Annals of Applied Statistics*, 8(3):1281 – 1313.

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

Fritsch, S., Guenther, F., and Wright, M. N. (2019). *neuralnet: Training of Neural Networks*. R package version 1.44.2.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

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

Grillenzoni, C. and Carraro, E. (2021). Sequential tests of causality between environmental time series: With application to the global warming theory. *Environmetrics*, 32(1):e2646.

Hadj-Amar, B., Finkenstadt, B., Fiecas, M., and Huckstepp, R. (2021). Identifying the recurrence of sleep apnea using a harmonic hidden Markov model. *Annals of Applied Statistics (In press)*.

Hooper, P. M. (1993). Iterative weighted least squares estimation in heteroscedastic linear models. *Journal of the American Statistical Association*, 88(421):179–184.

Hyndman, R. J. and Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, 26(3):1–22.

Jordan, A., Krueger, F., and Lerch, S. (2017). Evaluating probabilistic forecasts with the r package scoringrules.

Keenan, D. M. (1985). A tukey nonadditivity-type test for time series nonlinearity. *Biometrika*, 72(1):39–44.

Kleissl, J. (2013). *Solar Energy Forecasting and Resource Assessment*. Academic Press.

Magnus, J. R., Melenberg, B., and Muris, C. (2011). Global warming and local dimming: The statistical evidence. *Journal of the American Statistical Association*, 106(494):452–464.

Munkhammar, J., van der Meer, D., and Widén, J. (2019). Probabilistic forecasting of high-resolution clear-sky index time-series using a markov-chain mixture distribution model. *Solar Energy*, 184:688 – 695.

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

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

O'Hagan, A. and Forster, J. J. (2004). *Kendall's Advanced Theory of Statistics*, volume 2B of *Bayesian Inference*. Arnold, second edition.

Peters, A. and Hothorn, T. (2019). *ipred: Improved Predictors*. R package version 0.9-9.

REN21 (2020). Renewables 2020 Global Status Report.

Samanta, S., Prajneshu, and Ghosh, H. (2011). Modelling and forecasting cyclical fish landings: Setarma nonlinear time-series approach. *Indian Journal of Fisheries*, 58(3):39–43.

Sengupta, M., Xie, Y., Lopez, A., Habte, A., Maclaurin, G., and Shelby, J. (2018). The national solar radiation data base (NSRDB). *Renewable and Sustainable Energy Reviews*, 89:51 – 60.

Shepero, M., Munkhammar, J., and Widén, J. (2019). A generative hidden markov model of the clear-sky index. *Journal of Renewable and Sustainable Energy*, 11(4):043703.

Silva, I., Silva, M. E., Pereira, I., and Silva, N. (2005). Replicated INAR(1) processes. *Methodology and Computing in Applied Probability*, 7(4):517–542.

Tong, H. (2010). Threshold models in time series analysis—30 years on. *Statistics and Its Interface*, 2.

Tong, H. and Lim, K. S. (1980). Threshold autoregression, limit cycles and cyclical data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(3):245–292.

Tsay, R. S. (1986). Nonlinearity tests for time series. *Biometrika*, 73(2):461–466.

Voyant, C., Notton, G., Duchaud, J.-L., Almorox, J., and Yaseen, Z. M. (2020). Solar irradiation prediction intervals based on box–cox transformation and univariate representation of periodic autoregressive model. *Renewable Energy Focus*, 33:43 – 53.

Watier, L. and Richardson, S. (1995). Modelling of an epidemiological time series by a threshold autoregressive model. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 44(3):353–364.

Zhang, W., Kleiber, W., Florita, A. R., Hodge, B., and Mather, B. (2019). Modeling and simulation of high-frequency solar irradiance. *IEEE Journal of Photovoltaics*, 9(1):124–131.

Zhang, W., Kleiber, W., Hodge, B.-M., and Mather, B. (2021). A nonstationary and non-gaussian moving average model for solar irradiance. *Environmetrics*, page e2712.