Received: 20 March 2023 | Revised: 23 April 2023 | Accepted: 3 May 2023
DOI: 10.1002/sta4.583

ORIGINAL ARTICLE

WILEY

A space-time model with temporal cyclostationarity for probabilistic forecasting and simulation of solar irradiance data

Soumya Das[1] | Yasser M. Alshehri[2] | Georgiy L. Stenchikov[2] | Marc G. Genton[1]

[1]Statistics Program, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

[2]Earth Science and Engineering Program, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

Correspondence
Soumya Das, Statistics Program, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia.
Email: sundew122436@gmail.com

To establish a more sustainable future, Saudi Arabia is striving to reduce its dependency on fossil fuels and promote renewable energies. Solar energy is a major resource in Saudi Arabia because of the country's geographical location with year-round clear skies and plentiful sunlight. However, although solar energy is a clean and safe renewable energy resource, it can be quite unpredictable. Therefore, solar irradiance needs to be forecasted and simulated as accurately as possible on both regional and national scales in the Kingdom to assist in planning for reserve usage, switching sources and short-term power purchases. Based on an hourly solar diffuse horizontal irradiance (DHI) dataset from 45 different Saudi Arabian solar monitoring stations, this study proposes a novel spatio-temporal model. We identify the temporal dependency of DHI as cyclostationary and incorporate this key observation in the model. This feature helps yield accurate (i.e. significantly close to the ground truth and with narrow confidence bands) probabilistic forecasts and realistic simulations in near-future time points and at new locations. Then, the proposed model is compared with a stationary model that fails to recognize the cyclostationarity structure in the data. The proposed model produces significantly better predictions and simulations than the stationary model. Further, we calculate the photovoltaic power outputs using the simulated samples from both the models and the original observations and then compare them. The simulated samples from the proposed model afford photovoltaic power output estimates that are significantly closer to the original observed data than those from the stationary model, and therefore, they can better assist photovoltaic power output operators in assessing solar energy production and operational pre-planning to mitigate impacts by the uncertain nature of solar irradiance.

KEYWORDS
cyclostationary, energy production, periodic autoregressive and moving average, periodically correlated, prediction, renewable energy

1 | INTRODUCTION

Over the last few years, Saudi Arabia has been working to create a renewable energy sector with sufficient harnessing capabilities in terms of energy measurement, research and production. Moreover, in line with the *Saudi Vision 2030*, they have lately set a goal to generate 50% of the nation's power needs using renewable energy by 2030 during the 11th session of the General Assembly of the International Renewable Energy Agency (IRENA), in a speech by the president of King Abdullah City for Atomic and Renewable Energy (K.A.CARE). Among the renewable energy resources, solar energy is naturally a promising supply of sustainable power generation because of the geographically strategic location of

Saudi Arabia. For instance, a Saudi solar energy firm called "Desert Technologies" in cooperation with the French company "Green Corp Konnection" recently set up two solar energy containers named "Sahara," which can generate a large amount of power ($\sim$62 kW) to partially operate NEOM's assembly area for rally drivers, as published by Arab News on 17 January 2021. NEOM is a planned cross-border city in the Tabuk Province of northwestern Saudi Arabia that is designed to incorporate "smart city technologies" and is completely powered using renewable energy resources. This move has certainly set the stage for intensifying the solar energy provisions across Saudi Arabia, but it also owes cost effectiveness, performance accuracy and policy incentives.

Therefore, it is desirable to have solar irradiance predictors and simulators that are accurate, easily implementable and cost efficient on regional and national scales in Saudi Arabia. In addition, it is also important that the underlying idea of constructing the predictor or simulator can directly be applied to other countries' solar irradiance data with a slight or no modification. Indeed, not only Saudi Arabia but also many other countries around the globe have been positively affected by the surge of solar energy research and implementation. For instance, the United States (USA) has witnessed an increase of grid-connected solar photovoltaic (PV) capacity (i.e. the amount of power generated by a grid-connected solar PV system) by a factor of more than 109 from 2005 to 2015; see Wiser and Bolinger (2011) and Beiter et al. (2017). As another example, the City of Sydney, which is the central business district of Sydney, Australia, is now entirely powered by renewable energy, a substantial proportion of which is generated by two solar farms across regional New South Wales. Mostly, the use of PV power generators has increased manifold because of their easy installation, noiseless operation, zero greenhouse gas emission and fast return on investment; see Taghezouit et al. (2021). Consequently, numerous developments and research have been conducted on various solar resource measurements of various countries. Reikard (2009) employed the Box–Jenkins seasonal (or nonseasonal) autoregressive integrated moving average model (Box et al., 2015) to provide short-term forecasts using the data from various sites across the world. Pedro and Coimbra (2012) showed that the short-term forecasting efficacy of Artificial Neural Networks is comparable with that of the models in Reikard (2009). Ranganai and Sigauke (2020) used long memory models, and their hybrid versions to forecast irradiance 24 hours ahead at three different sites in South Africa; see also Tsekouras and Koutsoyiannis (2014), Lauret et al. (2017), Voyant et al. (2017) and Alparslan et al. (2019). Moreover, Lonij et al. (2013), Filik et al. (2017) and Li and Zhang (2020) forecasted power outputs using PV systems. Yang and van der Meer (2021) extensively reviewed post-processing in solar forecasting. Additionally, Myers et al. (1999), Pazheri (2014), Zell et al. (2015), AlYahya and Irfan (2016), and Salam and Khan (2018) studied different solar resources mainly for Saudi Arabia.

Acknowledging the above studies, herein, we present a novel spatio-temporal model that can produce high-quality (probabilistic) forecasts and realistic simulated solar irradiance data (specifically, diffuse horizontal irradiance or DHI) in space and time in order to help predict the PV power output both temporally and spatially. The identification of the periodic correlation structure (or cyclostationarity) of solar irradiance plays a pivotal role in the model's performance; see Hurd and Miamee (2007) and Napolitano (2019) for an introduction to cyclostationary processes. A second-order stochastic process $\{X_t, t \in \mathbb{Z}\}$ is called cyclostationary (or periodically correlated) with period T if its mean and covariance have periodic structures with a period of T. That is, for every $s, t \in \mathbb{Z}$,

$$m(t) = \mathrm{E}(X_t) = m(t+T) \text{ and } R(s,t) = \mathrm{E}[\{X_s - m(s)\}\{X_t - m(t)\}] = R(s+T, t+T) \tag{1}$$

holds, where T is the smallest positive integer that satisfies the above two identities.

In Section 3.1, we investigate and find out the common cyclostationarity structure of solar irradiance using data collected from three different representative locations: Haql (in Saudi Arabia), Colorado (in the USA) and Sydney (in Australia). In that subsection, we discover that, in general, solar irradiance exhibits a periodic pattern in not only its mean structure (which is well expected because of the Sun's periodic behaviour) but also its correlation structure. This can be understood intuitively. In general, the solar irradiance measured in the early hours of any day is statistically different from that measured during late hours. However, most probably, the solar irradiance measured at the same hour on different days is almost identical or closely correlated. Such behaviour can be indicative of an underlying cyclostationary process; see Napolitano (2019). Hurd and Miamee (2007) considered a similar example of hourly solar radiation levels taken at the meteorological station DELTA on Ellsmere Island, N.W.T., Canada, and showed that the data are cyclostationary.

Upon investigating the general behaviour of solar irradiance, based on this broader perspective, we propose a model for Saudi Arabia that can be easily extended to other parts of the world. For this purpose, we consider a DHI dataset measured at 45 different solar monitoring stations of Saudi Arabia, provided by K.A.CARE. We fit a purely cyclostationary spatio-temporal model (i.e. we do not consider any other covariate to explain the DHI). While implementing the proposed model for other countries, if a covariate is present for which the data are available and that significantly affects the DHI in that region, it can be incorporated in the model to improve it. However, from the above discussion and that in Section 3.1, one can still expect a cyclostationary process to explain the basic structure of the data quite well. Apart from using a cyclostationary process to model the temporal dynamics, we consider a Gaussian random field to capture the spatial dependencies. After proposing our model, we provide an estimation procedure and conduct spatial and spatio-temporal kriging. The estimation technique does not use popular approaches such as Bayesian or maximum likelihood (mainly to reduce the computational cost); contrastingly, our approach depends on the basic understanding of the data distribution across Saudi Arabia. We show that even this simple estimation technique can fit the model well and yields reliable kriging results. Thereafter, we compare the proposed model with a stationary one that overlooks the temporal periodic correlation structure in the data, as is the case of existing models to date. There, we establish a superior performance of the proposed model than the other. Additionally, we simulate the irradiance at a new location using both the models and again show the more realistic simulations produced by our model than those by the stationary model. Finally, using the simulated data, we calculate the power outputs for a standard PV power system, which is the applied goal of the predicted or simulated solar irradiances.

The detailed analysis of the space-time DHI data is conducted as follows. Section 2 introduces the concerned dataset and presents a preliminary analysis. In Section 3, the DHI data are identified as cyclostationary, and the model is presented. Section 4 describes a detailed inference procedure for the proposed spatio-temporal model. In Section 5, we discuss the inferential procedure to fit the model and provide spatial and spatio-temporal kriging. In Section 6, we assess the calibration of the probabilistic forecasts produced using the fitted model and we compare the proposed model with a classical stationary model that does not take account of the temporal periodic correlation structure. In Section 7, we simulate solar irradiance using both the models for further comparison. In Section 8, we calculate PV power generations. Finally, in Section 9, we summarize the study findings and discuss their significance.

2 | DHI DATA FROM SAUDI ARABIA

The analysed DHI dataset comprises hourly solar DHIs in watts per square metre (W/m^2) from 46 different stations dispersed over Saudi Arabia. To clarify, the data are observed and collected on the hour, not in the interval between 2 hours. The dataset is prepared by K.A.CARE as part of the Renewable Resource Monitoring and Mapping (RRMM) programme. It can be accessed via the following website (https://rratlas.kacare.gov.sa/). Additionally, the dataset contains global horizontal irradiance (GHI), direct normal irradiance (DNI), air temperature, humidity, wind speed, wind direction, zenith angle and azimuth angle data.

The DHI data at different stations were measured during different periods. For example, the data commence from 2013 for 32 stations, 2014 for 10 stations and 2015 for the remaining stations. For almost all of the stations, data recording ended in 2017, except for the Princess Norah University Station, where the data were last recorded on 1 December 2014. For the analysis, we consider the DHI data measured in a common interval, that is, from 00:00 AM on 1 September 2015 to 23:00 PM on 10 September 2015, for 45 stations; the data from the Princess Norah University Station were not considered. September was selected as it usually exhibits the highest DHI volatility, which was verified by plotting DHIs over different periods. Therefore, data from September will help establish the accuracy of the prediction capabilities of our proposed model at a new location and future time points. Further, as is known, DHI becomes zero (i.e. deterministic) at night. Therefore, for each day, we only consider the daytime hours (i.e. from 08:00 to 17:00). Although the Sun may set later than 17:00, the DHI observed during those later hours is quite low and, therefore, does not significantly contribute to the PV power outputs. Further, the censoring of the nighttime data is reasonable since we will be fitting a cyclostationary model, where the daytime hours are mostly affected by the daytime hours of the previous days.

The left panel of Figure 1 displays the locations of the 45 stations. Among the 45 stations, the 35 training stations, which are used for inference, are denoted by red triangles, and the remaining 10 testing stations, which are used for verification, are denoted by blue circles. We will

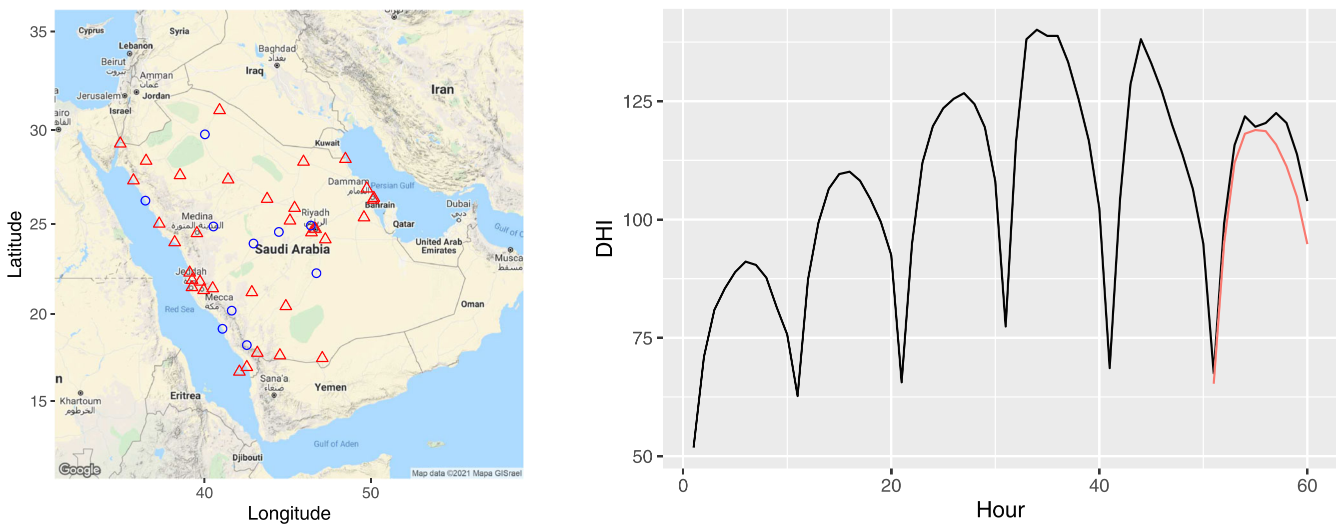

FIGURE 1 Left: Locations of the 45 solar monitoring stations in Saudi Arabia. Red triangles denote the 35 training stations, while blue circles denote the 10 testing stations. Right: Plot of the solar diffuse horizontal irradiance (DHI) at a representative station named *Saline water conversion corporation at Haql* observed during daytime hours (i.e. 08:00 to 17:00) from 1 September 2015 to 6 September 2015 (in black) and the predicted DHI during daytime hours on 6 September 2015 (in red). For the prediction, a PAR(1) model is used and trained using the daytime hours of the first 5 days.

discuss this in more detail in Section 5. The DHI data from a representative station, *saline water conversion corporation at Haql*, for the first 6 days of September 2015 are plotted in the right panel of Figure 1. The figure shows a recurring DHI behaviour with a period of 10, which is due to the Sun's periodic nature. Although we only observe the periodic nature of the mean function of the underlying process in the figure, the covariance function of the series also exhibits a periodic pattern (or cyclostationarity), as shown in Section 3.1.

Das et al. (2021) analysed a univariate solar GHI series, measured at King Abdullah University of Science and Technology (KAUST), Saudi Arabia, and provided a model for short-term temporal GHI forecasts. Whereas in this article, we extend this analysis to a spatio-temporal scenario. Although the temporal analysis in Das et al. (2021) is based on a GHI series, it can be easily reiterated for a DHI series with slight modifications. The modification lies in the consideration of the aerosol optical depth (AOD) as a covariate. For a GHI series, AOD can be incorporated in the model because of its strong correlation with GHI. In contrast, AOD does not exhibit a significant correlation with DHI and, therefore, may not be included in the model. For the Saudi Arabian stations other than KAUST, the AOD data are unavailable. Therefore, herein, we fit the DHI data using a purely cyclostationary model. Nevertheless, reliable forecasts and simulations are obtained using the proposed model.

Although we mainly focus on fabricating a model to yield spatial and spatio-temporal forecasts, for the sake of completeness, we briefly present a temporal cyclostationary model for a DHI series. We fit a periodic autoregressive moving average (PARMA) model to the first 5 days (i.e. 50 hours) of the DHI data measured at the Haql station and predict the data for the 6th day. For this, we denote the DHI time series as $\{y_t, t=1,\ldots,60\}$ and consider its first 50 data points to fit a PARMA(1,0) (or PAR(1)) model (the AR and MA orders of a PARMA model can be decided based on the periodic autocorrelation and partial autocorrelation plots):

$$y_t = \phi(t) y_{t-1} + \sigma(t)\epsilon_t, \; t = 1,\ldots,50, \tag{2}$$

where $\phi(t) = \phi(t+T)$ and $\sigma(t) = \sigma(t+T)$, for $t = 1,\ldots,50$, are model parameters, and $T = 10$ is the period of the series. To estimate the parameters of the series, we use the Yule–Walker estimation method; see Pagano (1978). Once fitted, we use the model to predict the next 10 hours (i.e. the daytime hours of the 6th day). The predicted values are plotted in red in the right panel of Figure 1. The figure clearly shows that a temporal cyclostationary model can fit the solar DHI data significantly well.

3 | SPATIO-TEMPORAL MODEL FOR DHI

In this section, we develop an appropriate spatio-temporal model for the solar DHI data.

3.1 | Temporal dynamics

As mentioned in Section 2, the DHI series measured at Haql station, Saudi Arabia, exhibits a cyclostationary pattern. Indeed, this behavior is not specific to Saudi Arabia. Rather, cyclostationarity seems to be a basic underlying structure of the solar DHI itself. To verify this, we consider the DHI data from three locations belonging to three different continents:

1. Central Sydney, Australia: The live data (i.e. real-time data of 7 days) are downloaded from the "Solcast API Toolkit," which can be accessed via https://toolkit.solcast.com.au/world-api. The dataset comprises $168 (= 24 \times 7)$ hourly solar DHIs collected from 18:00 on 8 March 2021 to 17:00 on 15 March 2021.
2. Denver, Colorado, USA: These data are obtained from the NREL-NSRDB website (https://maps.nrel.gov/nsrdb-viewer/). The considered dataset comprises a series of 168 hourly solar DHIs measured during the first 7 days of September 2015.
3. Saline water conversion corporation at Haql, Saudi Arabia: This is one of the 45 solar monitoring stations considered in this article. For the current purpose, we consider the hourly solar DHIs measured during the first seven days (i.e. the data comprise 168 data points) of September 2015.

Now, to check for cyclostationarity, we plot the sample squared coherence statistic for these three sets of solar DHI data. The sample squared coherence statistic at frequencies ω_p and ω_q is defined as (Hurd & Gerr, 1991):

$$|\mathcal{C}(p,q,M)|^2 = \frac{\left|\sum_{m=0}^{M-1} d_n(\omega_{p+m})\overline{d_n(\omega_{q+m})}\right|^2}{\sum_{m=0}^{M-1} |d_n(\omega_{p+m})|^2 \sum_{m=0}^{M-1} |d_n(\omega_{q+m})|^2},$$

where $\{d_n(\omega_k), k=0,\ldots,n-1\}$ are the coefficients of the discrete Fourier transform (DFT) of the corresponding time series, M denotes the smoothing parameter, $\omega_k = 2\pi k/n$ for $k=0,\ldots,n-1$ and n denotes the data size (i.e. $n=168$ in the current scenario). Hurd and Gerr (1991) mentioned that the DFT coefficient $d_n(\omega_k)$ approximates the increment of Z, an increment process of the concerned time series with its spectral measure defined as $\mathcal{F}((a,b]\times(c,d]) = \mathrm{E}[(Z(b)-Z(a))\overline{(Z(d)-Z(c))}]$ in the domain $(a,b]\times(c,d]$, in the kth interval $[\omega_k, \omega_{k+1})$. Here, the increment process $Z(\omega)$ and the corresponding time process $\{Y_t, t\in\mathbb{Z}\}$ are related as follows:

$$Y_t = \int_0^{2\pi} e^{it\omega} \mathrm{d}Z(\omega),\ t\in\mathbb{Z}.$$

Thus, the sample squared coherence statistic can be viewed as an "estimate" of $\mathrm{E}\left\{\mathrm{d}Z(\omega)\overline{\mathrm{d}Z(\lambda)}\right\}, (\omega,\lambda)\in[0,2\pi)^2$. This shall imply that, in a sample squared coherence plot, lines parallel to the main diagonal at distances $h=kn/T\,(k=0,\ldots,T-1)$ are an indication of the series being cyclostationary, where T denotes the period of the series; see also Martin and Kedem (1993), Dandawate and Giannakis (1994), Bloomfield et al. (1994), Broszkiewicz-Suwaj et al. (2003), Broszkiewicz-Suwaj et al. (2004), Wang et al. (2005) and Pries et al. (2018).

Figure 2 displays the sample squared coherence plots of the solar DHIs for the three considered locations. The left, middle and right panels of Figure 2 are representative of the data for Central Sydney, Denver and Haql, respectively. To calculate the sample squared coherence statistic, we consider the value of the smoothing parameter $M=32$ in all the three scenarios. The reciprocal of the smoothing parameter M is the spectral frequency resolution of the squared coherence estimate. A choice of M such that $n/M \gg 1$ often provides a reliable estimate. However, there are also some disadvantages of considering small values of M; see Hurd and Miamee (2007). The above choice of $M=32$ is reasonable since the value of $n=168$ is comparable with the data record lengths adopted in Hurd and Gerr (1991).

For the squared coherence plot, since we considered $n=168$ data points and the solar period is $T=24$, lines parallel to the main diagonal at every $h=168/24=7$ unit distance shall indicate the cyclostationary structure of the data, and this is true for the three cases presented in Figure 2. We note that the lines appear clearer for Haql than that of Sydney or Denver. This is justified since Saudi Arabia mostly experiences a stable and consistent weather pattern (i.e. few clouds and occasional dust storms) year round unlike other countries, and therefore, the periodicity is less influenced in Saudi Arabia than in Australia or the USA.

The above discussion suggests that a cyclostationary model is appropriate for fitting solar DHI data of not only Saudi Arabia but also any part of the globe at any time of the year. Thus, a cyclostationary process is fitted to the temporal part of the spatio-temporal model. Specifically, we fit a PAR(1) model, as defined in (2), to describe the temporal dynamics of the space-time solar DHI dataset of Saudi Arabia. Remarkably, there are stationary models, such as a seasonal autoregressive model, that can also explain the periodic behaviour in the mean and auto-covariance structure of the data. However, they fail to explain the periodic covariance structure of the data.

In the next subsection, we propose a spatio-temporal model. For convenience, we standardize the dataset for each of the 45 stations, that is, we subtract its mean and divide by its standard deviation for the remaining analysis. Hereafter, "DHI" will denote "standardized DHI" unless otherwise stated. This helps us to consider a simplified model without the intercept. Furthermore, unless standardized, the observed large values of DHIs might need extra care while analysing. Standardizing them will solve this issue. Nevertheless, note that the conclusions deduced from the following analysis are valid even if the data are not standardized.

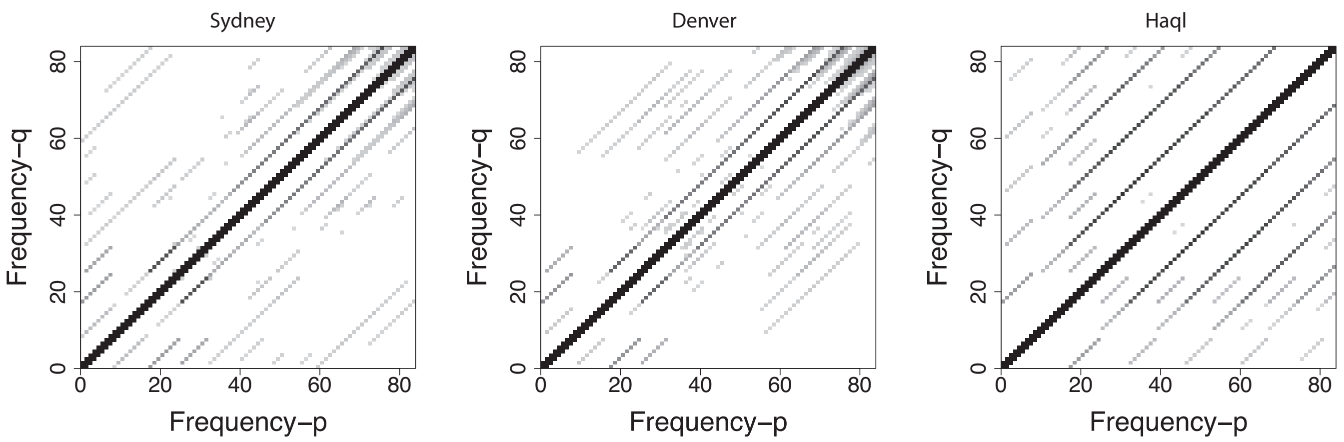

FIGURE 2 Squared coherence plots of the hourly solar diffuse horizontal irradiances (DHIs) collected over seven days at three different locations: Central Sydney in Australia (left), Denver in the USA (middle) and Haql in Saudi Arabia (right)

3.2 | Spatio-temporal model

Besides studying the temporal nature of the DHI, it is also essential to investigate its spatial structure. This is because given the observed irradiances at several stations across the nation, a spatio-temporal model will help predict the DHI at new locations. We propose a spatio-temporal model as follows.

Let $Z_t(\boldsymbol{s})$ denote the (standardized) DHI at location $\boldsymbol{s}$ and time point t, where $\boldsymbol{s} \in \{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_L\}$ and $t \in \{1,2,\ldots,n\}$. Here, $L =$ number of stations $=$ 45 and $n =$ number of regularly spaced time points at each station $=100$ (10 daytime hours of 10 consecutive days, i.e., from 08:00 to 17:00 on 1–10 September 2015). Then, we consider the following spatio-temporal (cyclostationary) model:

$$Z_t(\boldsymbol{s}) = Y_t(\boldsymbol{s}) + \epsilon_t(\boldsymbol{s}) \text{ with } Y_t(\boldsymbol{s}) = \phi(t) Y_{t-1}(\boldsymbol{s}) + \sigma_\eta(t)\eta_t(\boldsymbol{s}), \tag{3}$$

where $\{\epsilon_t(\boldsymbol{s})\}$ is a white-noise process which is assumed to independently follow a $\mathcal{N}(0,\sigma_\epsilon^2)$ distribution. That is, σ_ϵ^2 is the measurement error or the nugget effect. In the process model, $\{\eta_t(\boldsymbol{s})\}$ is an orthonormal sequence. The model parameters $\{\phi(t)\}$ and $\{\sigma_\eta(t)\}$ satisfy $\phi(t) = \phi(t+T)$ and $\sigma_\eta(t) = \sigma_{\eta(t+T)}$, respectively, for all $t = 1,\ldots,n$. Here, $T =$ the common period of every DHI series at each station $=10$. Further, the parameters $\{\phi(t)\}$ are required to satisfy $\left|\prod_{t=1}^{T} \phi(t)\right| < 1$, which assures that the cyclostationary process is bounded (Hurd & Miamee, 2007, Theorem 8.1 on p. 226). It is noteworthy that we have not included a purely spatial term in the model to allow for different overall means at different sites. This is reasonable in our context because we do not observe much spatial variations across regions in Saudi Arabia. For countries with significant spatial variations (for example, different altitudes), the model (3) can be modified to include a purely spatial term. For later use, we write the observed data in a matrix form of dimension $L \times n$ as $\boldsymbol{Z} = (\boldsymbol{Z}(\boldsymbol{s}_1)^\top,\ldots,\boldsymbol{Z}(\boldsymbol{s}_L)^\top)^\top$, where $\boldsymbol{Z}(\boldsymbol{s}_l) = (Z_1(\boldsymbol{s}_l),\ldots,Z_n(\boldsymbol{s}_l))^\top$.

To describe the spatial dependency, we assume the commonly used Gaussian random field (GRF) (see Cressie & Wikle, 2011) with mean zero, through the term $\eta_t(\boldsymbol{s})$. Moreover, we assume an isotropic exponential correlation function:

$$\operatorname{cov}(\eta_t(\boldsymbol{s}), \eta_{t'}(\boldsymbol{s}')) = \begin{cases} \exp(-\|\boldsymbol{s}-\boldsymbol{s}'\|/\beta), & \text{if } t = t', \\ 0, & \text{if } t \neq t', \end{cases}$$

for $\boldsymbol{s} \neq \boldsymbol{s}'$, where $\beta > 0$ is an unknown range parameter. The exponential correlation function is employed over the commonly used Matérn correlation function because of its simplicity (it contains one parameter less than the Matérn function for estimation), fewer considered locations (i.e. the smoothness parameter of the Matérn function is difficult to estimate with only 45 locations), and rough field of errors (which is verified via a preliminary analysis).

4 | INFERENCE

In this section, we discuss the fitting of the spatio-temporal model in detail. First, we note that although the spatio-temporal model is quite simple, it has too many parameters. In fact, the parameter set of interest contains 22 parameters:

$$\{\sigma_\epsilon^2, \phi(1),\ldots,\phi(T), \sigma_\eta^2(1),\ldots,\sigma_\eta^2(T), \beta\}.$$

This definitely creates an impediment (e.g. in terms of computational time) while using the standard estimation technologies, such as maximum likelihood estimation (MLE) or Bayesian methodologies. Therefore, we adopt a simple two-stage method as follows:

Stage 1: Since the weather is quite stable across Saudi Arabia, the estimates of the cyclostationary parameters, $\{\phi(1),\ldots,\phi(T),\sigma_\eta^2(1),\ldots,\sigma_\eta^2(T)\}$, that are found individually for each station will not significantly vary. Therefore, we propose simply estimating them for each of the 45 stations (using Yule-Walker estimation method; see Pagano [1978]) and plugging-in the averages of the corresponding parameter estimates in the model; see Basawa and Lund (2001) for properties of PARMA parameter estimates.

Although this proposal may not be in line with the conventional method of estimating the model parameters, since the empirical estimate of the nugget effect can be verified to be quite small, especially for the concerned dataset, this method works and consequently saves considerable computational time.

For a country, where the uniform nature of the parameters over space cannot be guaranteed (e.g., a country with a considerable number of mountains, hills, plateaus, and plains and, therefore, an unequal distribution of solar irradiance), the standard MLE or Bayesian method can be used. Alternatively, each of the cyclostationary parameters can be assumed to be spatially varying. Then, for each of the 20 parameters, spatial kriging can be performed to determine parameter estimates at a new location (instead of just averaging them over the stations).

Stage 2: Once we plug-in the cyclostationary parameters in the spatio-temporal model, we estimate the remaining two parameters, σ_ϵ^2 and β, using MLE (see Cressie & Wikle, 2011).

For this, we need to derive the covariance matrix, $\boldsymbol{\Sigma}_Y$, of

$$\boldsymbol{Y} = (\boldsymbol{Y}(\boldsymbol{s}_1)^\top, \ldots, \boldsymbol{Y}(\boldsymbol{s}_L)^\top)^\top,$$

where $\boldsymbol{Y}(\boldsymbol{s}_l) = (Y_1(\boldsymbol{s}_l), \ldots, Y_n(\boldsymbol{s}_l))^\top$. Here, we consider a separable covariance model, $\boldsymbol{\Sigma}_Y = \boldsymbol{\Sigma}_L \otimes \boldsymbol{\Sigma}_n$, where $\boldsymbol{\Sigma}_L$ and $\boldsymbol{\Sigma}_n$ are $L \times L$ and $n \times n$ dimensional matrices, respectively, and $\otimes$ denotes the Kronecker product (Van Loan, 2000). The assumption of separability reduces a substantial amount of computational time. Simultaneously, in many cases, this assumption does not yield significant differences in prediction errors; see Genton (2007). Therefore, with a slight compromise of accuracy, the separability of the covariance matrix seems reasonable to assume. However, in some scenarios, the separability assumption may not afford satisfactory results. Then, an alternative approach could be to either approximate the likelihood function or use *ExaGeoStat*, which is a high performance software for large-scale geostatistics in climate and environment modeling (Abdulah et al., 2018), to determine the exact MLE.

The construction of $\boldsymbol{\Sigma}_L$ is straightforward. That is, the (i,j)th element of the matrix is

$$\boldsymbol{\Sigma}_L^{ij} = \exp(-\|\boldsymbol{s}_i - \boldsymbol{s}_j\|/\beta),$$

for all $i,j \in \{1,\ldots,L\}$. To find the matrix $\boldsymbol{\Sigma}_n$, we need the expressions of $\boldsymbol{\Sigma}_n^{tt'} = \mathrm{E}(Y_t(\boldsymbol{s})Y_{t'}(\boldsymbol{s}))$ for all $t,t' \in \{1,\ldots,n\}$ and any representative $\boldsymbol{s} \in \{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_L\}$ (note that the process was standardized to a zero mean). The derivation of this covariance function is similar to Hurd and Miamee (2007). For the sake of completeness, we mention the derived final expressions. Let

$$B_0(t) = 1 \text{ and } B_k(t) = \prod_{j=0}^{k-1} \phi(t-j),\ k = 1,\ldots,T$$

(Hurd & Miamee, 2007, p. 26). Then, for $t' > t$,

$$\begin{aligned} \mathrm{E}(Y_t(\boldsymbol{s})Y_{t'}(\boldsymbol{s})) &= \sigma_Y^2(t) \prod_{j=t+1}^{t'} \phi(j) \text{with} \\ \sigma_Y^2(t) = \mathrm{E}(Y_t(\boldsymbol{s})^2) &= \frac{1}{1 - B_T^2(t)} \sum_{j=0}^{T-1} B_j^2(t)\sigma_\eta^2(t-j). \end{aligned}$$

To visualize the temporal covariance matrix $\boldsymbol{\Sigma}_n$, we plot its heatmap in Figure 3. The heatmap is constructed using the parameter estimates based on the 35 training stations. We observe diagonally arranged blocks in the figure since the covariance function is periodic.

After we calculate the covariance matrix $\boldsymbol{\Sigma}_Y = \boldsymbol{\Sigma}_L \otimes \boldsymbol{\Sigma}_n$, we optimize the Gaussian likelihood with respect to the remaining two parameters, σ_ϵ^2 and β:

$$\mathcal{L}(\sigma_\epsilon^2, \beta | \boldsymbol{Z}) \propto |\boldsymbol{\Sigma}_Z|^{-1/2} \exp\left(-\frac{1}{2}\boldsymbol{Z}^\top \boldsymbol{\Sigma}_Z^{-1}\boldsymbol{Z}\right), \tag{4}$$

where $\boldsymbol{\Sigma}_Z = \boldsymbol{\Sigma}_Y + \sigma_\epsilon^2 \boldsymbol{I}$ and $\boldsymbol{I}$ denotes the identity matrix of a suitable dimension.

5 | IMPLEMENTATION AND PREDICTION

5.1 | Spatial prediction

To implement the inference scheme described in the previous section, the dataset is randomly split into two sets: a training set containing 100 hours (i.e. from 08:00 to 17:00 on September 1-10, 2015) of solar DHIs from 35 stations (say, $\boldsymbol{s}_1,\ldots,\boldsymbol{s}_{35}$) and a testing set containing the same 100 hours for the remaining 10 stations (say, $\boldsymbol{s}_{36},\ldots,\boldsymbol{s}_{45}$). The training (red triangle) and testing (blue circle) stations are presented in the left panel of Figure 1.

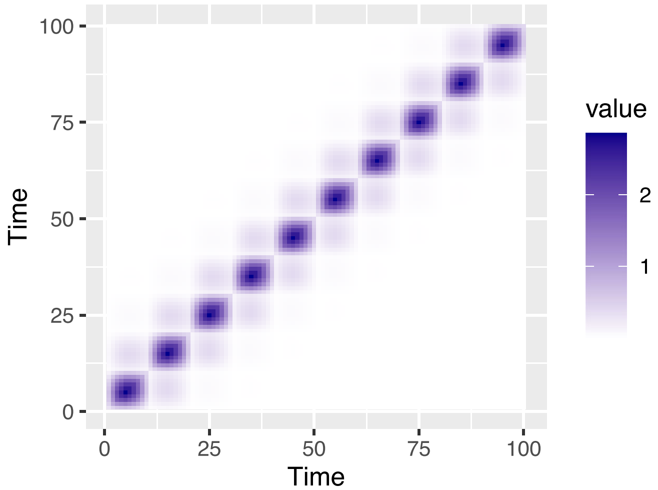

FIGURE 3 Heatmap of the covariance matrix $\boldsymbol{\Sigma}_n$ for the Saudi diffuse horizontal irradiance (DHI) data.

TABLE 1 Parameter estimates of the cyclostationary model parameters using the two-step estimation technique discussed in Section 4. The estimates were determined using the DHI data of the 35 training stations.

	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$	$t=6$	$t=7$	$t=8$	$t=9$	$t=10$
$\hat{\phi}(t)$	0.280	1.538	1.300	1.093	1.049	0.914	0.869	0.785	0.679	0.501
$\hat{\sigma}_\eta^2(t)$	0.262	0.205	0.266	0.257	0.224	0.221	0.237	0.209	0.206	0.183

To estimate the cyclostationary model parameters for each of the 35 stations, we use the Yule–Walker estimation method; see Pagano (1978). The estimates of the model parameters of (3) (averaged Yule-Walker estimates of 35 stations),

$$\{\phi(1),\ldots,\phi(T),\sigma_\eta^2(1),\ldots,\sigma_\eta^2(T)\},$$

are tabulated in Table 1. We observe that the parameter estimates $\{\hat{\sigma}_\eta^2(t), t=1,\ldots,T\}$ have less variance. However, this dispersion seems significant considering that we have standardized the data prior to analysis. Therefore, it may not suggest $\{\sigma_\eta^2(t)\}$ constant over time. Finally, we plug-in these estimates to the model and optimize (4) to determine the nugget and the range parameter estimates as $\hat{\sigma}_\epsilon^2=0.022$ and $\hat{\beta}=325.135$, respectively, the latter indicating a strong spatial dependence structure.

After the model parameter estimation, we provide point predictions of the solar DHIs at the 10 test stations, which were not used during the model training. For the prediction, we employ the standard kriging technology mentioned in Cressie (1993). In other words, the predicted values at the test stations are simply the conditional mean (under Gaussian assumption) of the DHIs at the test stations, given the observed data at the training stations, $\hat{\mathbf{Z}}_{\text{test}}=\mathrm{E}(\mathbf{Y}_{\text{test}}|\mathbf{Z}_{\text{train}})$, where $\mathbf{Y}_{\text{test}}=(\mathbf{Y}(\mathbf{s}_{36})^\top,\ldots,\mathbf{Y}(\mathbf{s}_{45})^\top)^\top$ and $\mathbf{Z}_{\text{train}}=(\mathbf{Z}(\mathbf{s}_1)^\top,\ldots,\mathbf{Z}(\mathbf{s}_{35})^\top)^\top$. The expression of this conditional expectation is derived in Cressie (1993) (p. 828). However, for convenience, we state its final expression as follows. Let $\mathbf{Y}_{\text{train}}=(\mathbf{Y}(\mathbf{s}_1)^\top,\ldots,\mathbf{Y}(\mathbf{s}_{35})^\top)^\top$ and

$$\begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{21}^\top \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

be the covariance matrix of $(\mathbf{Y}_{\text{test}}^\top,\mathbf{Y}_{\text{train}}^\top)^\top$, that is, $\mathrm{cov}(\mathbf{Y}_{\text{test}})=\boldsymbol{\Sigma}_{11}$, $\mathrm{cov}(\mathbf{Y}_{\text{train}},\mathbf{Y}_{\text{test}})=\boldsymbol{\Sigma}_{21}$ and $\mathrm{cov}(\mathbf{Y}_{\text{train}})=\boldsymbol{\Sigma}_{22}$. Then,

$$\mathrm{E}(\mathbf{Y}_{\text{test}}|\mathbf{Z}_{\text{train}})=\boldsymbol{\Sigma}_{21}^\top(\hat{\sigma}_\epsilon^2\mathbf{I}+\boldsymbol{\Sigma}_{22})^{-1}\mathbf{Z}_{\text{train}}. \tag{5}$$

Furthermore, we state the expression for the conditional covariance that is required to determine the prediction uncertainty (or predictive interval):

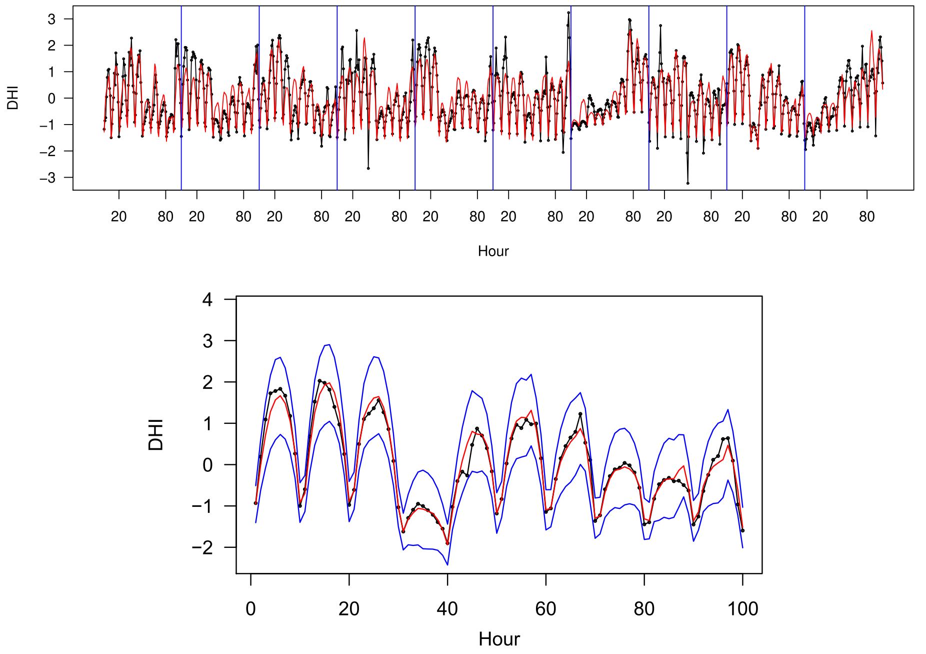

FIGURE 4 Top: Predicted 10 days of the 10 test stations (red) superimposed on the originally observed data at those stations (black). The 10 stations are serially appended and vertical blue lines separate them. Bottom: Predicted 10 days at Station 9 (red) and the corresponding 95% confidence band (blue) superimposed on the originally observed data at that station (black)

$$\mathrm{cov}(\boldsymbol{Y}_{\text{test}}|\boldsymbol{Z}_{\text{train}}) = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{21}^{\top}\left(\hat{\sigma}_{\epsilon}^{2}\boldsymbol{I} + \boldsymbol{\Sigma}_{22}\right)^{-1}\boldsymbol{\Sigma}_{21}. \tag{6}$$

The top panel of Figure 4 presents the predicted values (in red) of the 10 days at the 10 test stations and the originally observed values (in black). For better visualization, all 10 stations are serially appended in the same figure with vertical blue lines separating them. The figure shows that the predicted curve well approximates the observed curve. For example, we notice that although extreme solar DHI values are experienced on the fourth day at Station 9, the red curve still well approximates them. This suggests that the inference scheme presented in the preceding section, although seemingly quite simple, performs satisfactorily. Moreover, we calculate the 95% confidence band of these predicted values using (5) and (6). We do not overlap the confidence band with the plot in the top panel of Figure 4 to keep it intelligible. Rather, in the bottom panel of Figure 4, a separate plot is presented for one representative station, Station 9, wherein the confidence band is in blue. We observe that the originally observed values (in black) always stay within the confidence band.

5.2 | Spatio-temporal prediction

Since the proposed model can produce reasonable predictions at new spatial locations, we further investigate the forecasts at new stations and future time points. Therefore, we consider the first 90 hours of the concerned time range from each of the 35 training stations to fit the model and predict the last 10 hours of the 10 test stations. This implementation scheme is similar to that in the previous subsection. Figure 5 plots the predicted values (in red) of the last 10 hours for four representative stations. The original values are plotted in black, and the corresponding 95% confidence bands are in blue. Similar to the spatial case, we note the good prediction quality of the proposed model in this spatio-temporal scenario.

6 | VERIFICATION, EVALUATION AND COMPARISON

6.1 | Verification

To verify probabilistic forecasts, one of the most popular tools is implementing a calibration-based framework of these forecasts. To put in simple terms, if a well-calibrated model suggests that the probability of occurrence of an event is $\alpha\%$, then the observed frequency of that event

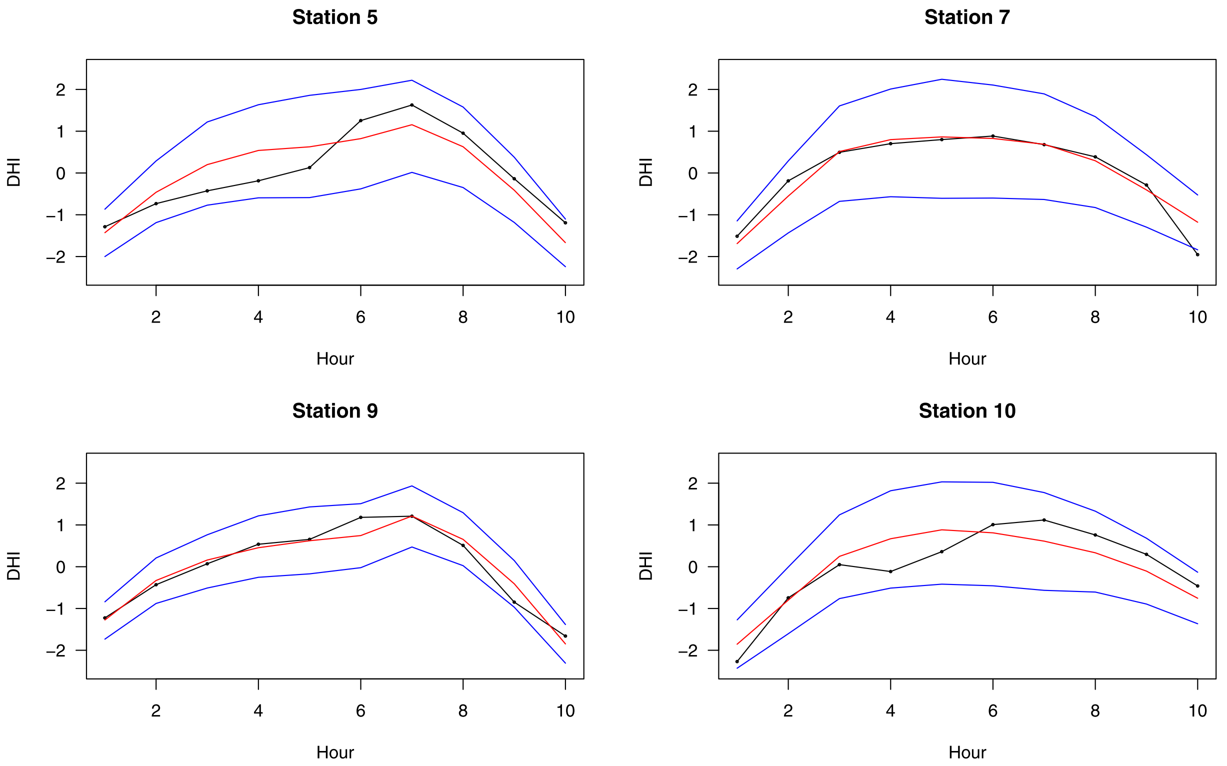

FIGURE 5 Predicted last 10 hours at Stations 5, 7, 9 and 10 (red) and the corresponding 95% confidence band (blue) superimposed on the originally observed data at those stations (black)

occurring should also be, on average, $\alpha\%$. In other words, we need to compare the empirical coverage rates (or observed frequencies) with the nominal coverage rates to verify the calibration of the forecasts. A popular approach is to plot the reliability diagram (i.e. empirical coverage rate versus nominal coverage rate); see Pinson and Kariniotakis (2010). In the reliability diagram plot, the closer the empirical coverage rates are to the nominal coverage rates (i.e. the diagonal line), the better the calibration is. Below, we calculate the empirical coverage rates for our current scenario.

Let $q_t^{\alpha}(\mathbf{s}_l)$ denote the quantile forecast at time point t and location $\mathbf{s}_l$. Then, we define an indicator variable $I_t^{\alpha}(\mathbf{s}_l)$ as follows:

$$I_t^{\alpha}(\mathbf{s}_l) = \begin{cases} 1, & \text{if} Z_t(\mathbf{s}_l) \leq q_t^{\alpha}(\mathbf{s}_l), \\ 0, & \text{otherwise.} \end{cases}$$

That is, the indicator variable assigns either 1 (hit) or 0 (miss) to each of the observed values. We further define the sum of hits as follows:

$$h^{\alpha} = \sum_{t \in \mathbb{D}_1} \sum_{l \in \mathbb{D}_2} I_t^{\alpha}(\mathbf{s}_l),$$

where $\mathbb{D}_1$ and $\mathbb{D}_2$ denote the sets of considered time points and locations, respectively. Then, the empirical coverage rate is calculated as

$$c^{\alpha} = \frac{h^{\alpha}}{|\mathbb{D}_1||\mathbb{D}_2|},$$

where $|\cdot|$ denotes the cardinality of a set.

To calculate c^{α} numerically, we consider the setting described in Section 5.2. Figure 6 presents the diagram corresponding to the spatio-temporal forecasts obtained using the proposed model. In the figure, the empirical coverage rates (in red) almost overlap with the diagonal (i.e. the nominal coverage rate), suggesting a significantly good forecast calibration.

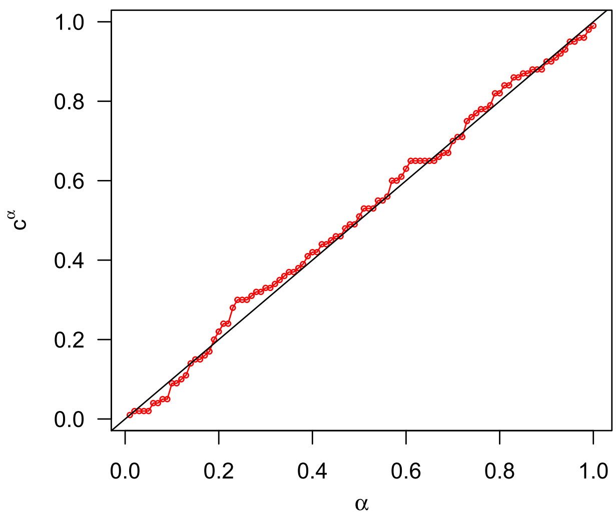

FIGURE 6 Reliability diagram of the spatio-temporal forecasts obtained in Section 5.2 using the newly proposed model.

6.2 | Evaluation and comparison

In this section, to evaluate our cyclostationary (i.e. PAR) spatio-temporal model, we compare it with another model that does not recognize the cyclostationary structure in the covariance function of the DHI series. To be specific, we compare our model with a stationary isotropic model where the PAR process model is replaced with an autoregressive (AR) process model of order one, that is, the model parameters $\{\phi(t)\}$ and $\{\sigma_\eta(t)\}$ are assumed to be constant for all $t = 1,...,T$. In other words, we compare the difference in prediction performance between constant and periodic covariance structures. To estimate the model parameters of the stationary model, we employ a similar but accordingly modified inference scheme as that described in Section 4.

To evaluate and compare the probabilistic forecasts produced by our model and the stationary model, we implement the widely used diagnostic tool, probability integral transform (PIT) plots; see Dawid (1984). For a time point t and location $\boldsymbol{s}_l$, it is defined as

$$\mathrm{PIT}_t(\boldsymbol{s}_l) = \int_0^{Z_t(\boldsymbol{s}_l)} f(u)\mathrm{d}u,$$

where f is the probability density of the underlying solar DHI distribution; the function f is assumed to be continuous in our scenario. Dawid (1984) proved that the PIT values are uniformly distributed on $[0,1]$ whenever the forecasts are ideal. In a practical scenario, the closer the PIT histogram is to a $\mathrm{U}[0,1]$, the better the probabilistic forecasts are. To quantify this goodness of fit, we calculate the Euclidean distance between the histogram and the horizontal line at one unit distance from the x-axis (in the first quadrant), which we call a square root sum of the squared distances (SRSSD).

To implement this method, we need the kriging outputs for each of the 45 stations using both models. Therefore, we conduct a spatial kriging for all 45 stations for both models using the "leave-one-station-out" technique. That is, for each of the 45 stations, we consider the other 44 stations to predict 10 days for that station.

Before calculating the PIT values and plotting them using these forecasts, we plot the predicted values produced by both the models and the observed data for one representative station (Haql) to visualize the prediction differences of the two models; see Figure 7. There, we plot the observed DHI data in Haql in black. The predicted values using the proposed and stationary models are superimposed in red and blue, respectively. The figure clearly shows that when the volatility is higher, the proposed model better captures the structure of the observed data than the stationary model.

Next, we calculate the PIT values and plot their histograms for the cyclostationary (left) and stationary (right) models in Figure 8. The figure shows that the PIT histogram of the proposed model is closer to a $\mathrm{U}[0,1]$ distribution (whose probability density function or pdf is marked in dotted red along $y = 1$) than that of the stationary model. Additionally, we calculate the respective SRSSD values. The SRSSD values for the cyclostationary and stationary models are 0.912 and 1.106, respectively. This further establishes the superiority of the proposed model.

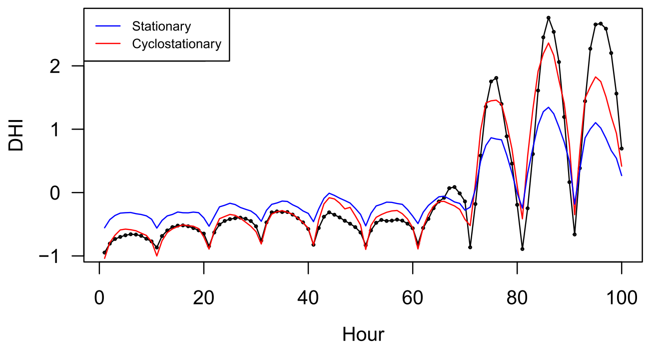

FIGURE 7 Predicted 10 days of solar diffuse horizontal irradiances (DHIs) at the solar monitoring station Haql using the proposed model (in red) and the stationary model (in blue). The observed values of the DHIs at the same station are plotted in black.

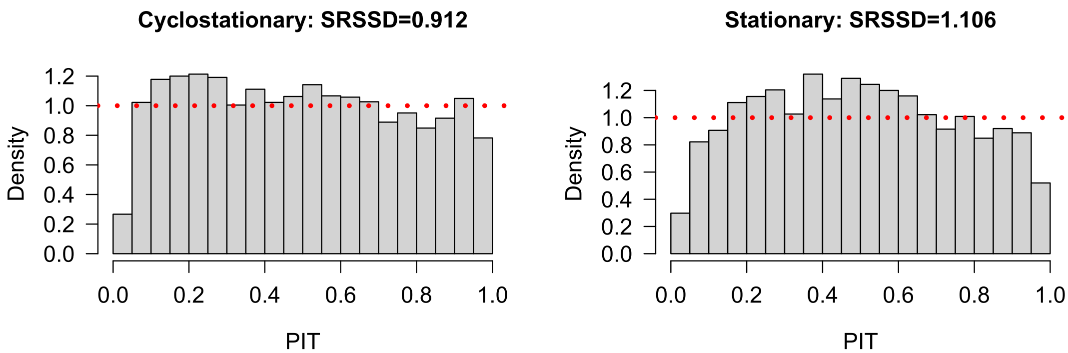

FIGURE 8 Probability integral transform (PIT) histogram plots of the forecasts using the proposed spatio-temporal cyclostationary model (left) and the stationary model, which does not consider the periodic correlation structure in the data (right). The dotted red lines at $y=1$ in the histograms indicate the pdf of U[0,1]. The corresponding square root sum of the squared distance (SRSSD) values are mentioned on top of each histogram.

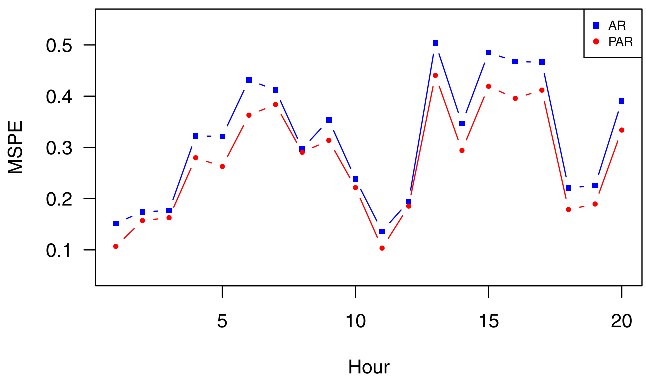

FIGURE 9 Plot of mean squared prediction error (MSPE), averaged over stations, of the diffuse horizontal irradiance (DHI) predictions using our model (red) and the stationary model (blue) for the last 2 days. In the legend, AR denotes the stationary model with an AR process model, while PAR denotes the proposed spatio-temporal model with a PAR process model.

To further compare these two models, we calculate the mean squared prediction errors (MSPE; see page 270 of Wikle et al. (2019)) of the corresponding predicted values for each station over the last two days (i.e. for 20 hours), average them over stations and plot them against time; see Figure 9. We observe that the cyclostationary model always has better MSPE than the stationary model.

Additionally, the cross-validation score (CVS) and the multivariate continuous ranked probability score (CRPS) or energy score (ES) can be employed to compare the models. CVS can be defined as

$$\mathrm{CVS}_{\mathrm{model}} = \frac{1}{45n}\sum_{l=1}^{45}\sum_{t=1}^{n}\left\{Z_t(\boldsymbol{s}_l) - \hat{Z}_t^{(-l)}(\boldsymbol{s}_l)\right\}^2,$$

where $\hat{Z}_t^{(-l)}(\boldsymbol{s}_l)$ indicates the predicted value of $Z_t(\boldsymbol{s}_l)$ using the data from all the stations except the station $\boldsymbol{s}_l$. The theoretical definition of ES and its detailed implementation process are presented in pages 277 and 300, respectively, of Wikle et al. (2019). Evidently, for both CVS and ES, the smaller the measure, the better the prediction. For example, for the PAR and AR process models, the CVS was 0.218 and 0.241, respectively, indicating that the proposed model performed 10% better than the stationary model.

For a comprehensive comparison between these two models, we calculate both scores (CVS and ES) for an increasing number of stations (i.e. 10, 15, 20, 25, 30, 35, 40 and 45) and plot them; see the supporting information. The plots show that, as the number of stations increases, both scores decrease. Moreover, similar to the previous diagnostics, the cyclostationary model always outperforms the stationary model.

7 | SIMULATION OF SOLAR DHI

Besides forecasting ahead in time or at a new spatial location, we can use the proposed model to simulate solar DHIs at a new location. Since DHI is a key factor in power calculation, simulating an entire series of solar DHIs at a new location is practically important.

Therefore, in this section, we simulate 10 DHI time series. Each series is simulated for 10 daytime hours of 10 consecutive days, that is, from 08:00 to 17:00 on 1–10 September 2015, at the Haql station. The model parameters are estimated using the data of the remaining 44 stations in the same time range and by considering both the cyclostationary model with PAR process parameters and the stationary model with AR process parameters.

Once the parameters are estimated, we calculate the conditional mean and covariance of the solar DHI at Haql given the observed data at the remaining stations; that is, $\mathrm{E}(\boldsymbol{Y}_{\mathrm{Haql}}|\boldsymbol{Z}^{(-\mathrm{Haql})})$ and $\mathrm{cov}(\boldsymbol{Y}_{\mathrm{Haql}}|\boldsymbol{Z}^{(-\mathrm{Haql})})$, under Gaussian assumption, where $\boldsymbol{Z}^{(-\mathrm{Haql})}$ denotes the entire observation vector with the data for the station Haql removed. The derivation of these two terms can be found in page 828 of Cressie (1993), and their final expressions are provided in Section 5. Then, we simulate 10 samples from a multivariate Gaussian distribution with mean $\mathrm{E}(\boldsymbol{Y}_{\mathrm{Haql}}|\boldsymbol{Z}^{(-\mathrm{Haql})})$ and covariance $\mathrm{cov}(\boldsymbol{Y}_{\mathrm{Haql}}|\boldsymbol{Z}^{(-\mathrm{Haql})})$.

Figure 10 presents the functional boxplots (Sun & Genton, 2011) of the 10 DHI series simulated using the cyclostationary (in the left panel) and stationary (in the right panel) models. The originally observed values during the 10 days of the time period are superimposed on both boxplots in green. From the plots, we can conclude that our proposed model produces more realistic simulations than the stationary model.

8 | ENERGY PRODUCTION

Finally, one of the main objectives of forecasting and simulating solar DHI is to directly use the forecasts and simulated samples in power calculation because this will assist PV power output operators in operational pre-planning, purchasing short-term power resources and taking early measures to mitigate impacts by the uncertain nature of solar irradiance, such as introducing new market products and increasing the procurement of

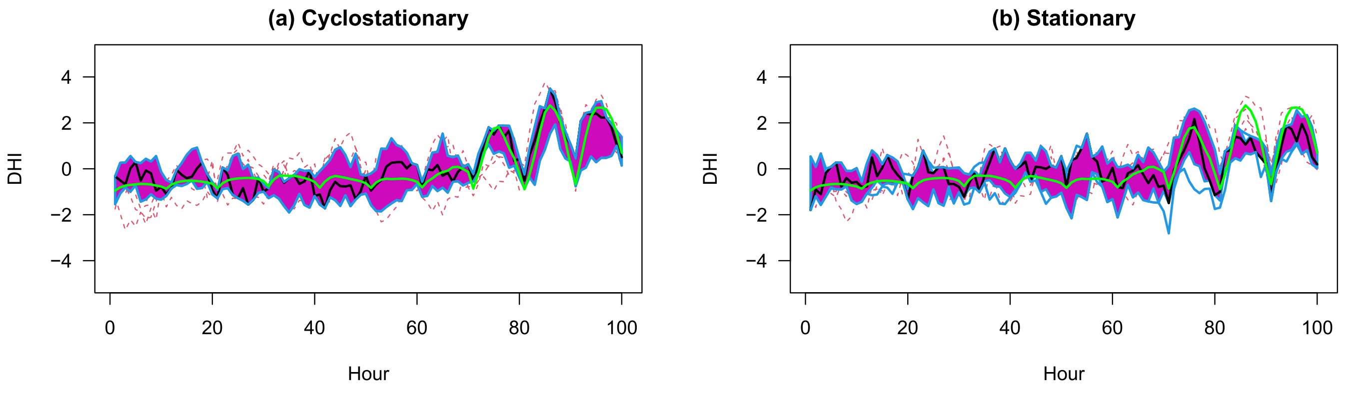

FIGURE 10 Functional boxplots of the 10 simulated series of 10 days at the solar monitoring station Haql using (a) the proposed model with PAR process parameters and (b) the stationary model with AR process parameters. The originally observed data during that time period are plotted in green.

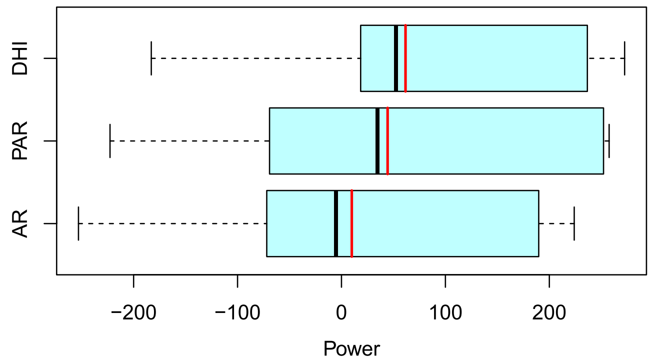

FIGURE 11 Boxplots of the average power calculated using 10 days of the observed data at the Haql station (top), 10 days of simulated series at the same station using our proposed model with PAR process parameters (middle), and the stationary model with AR process parameters (bottom). The red and black lines in each boxplot indicate the mean and median, respectively, of the corresponding series.

operating reserves. In this section, we establish that the estimated power output using the simulated DHI samples from our model is more reliable than that produced by the stationary model. First, we simulate solar DHI data at the Haql station for a period of 10 days using both cyclostationary and stationary models and consider the 10 days of originally observed data in the same time range as that in the previous section. Then, to estimate the average power generation on a typical day, we calculate the direct current (DC) power outputs by a PV power system for all the 100 daytime hours of the 10 days and average them on an hourly basis (i.e. we average the calculated power outputs at 08:00 each day, 09:00 each day, and so on) for each of these three solar DHI series. Standard PV module specifications are used (see Taghezouit et al., 2021) and therefore, this setting can be installed in Saudi Arabia to utilize the following results.

We note that, to calculate the DC power outputs, we also require GHI, DNI, air temperature, zenith angle and azimuth angle data from the Haql station, which are available in the K.A.CARE dataset, as mentioned in Section 2. To compute the DC power output, the PVWatts module model by the National Renewable Energy Laboratory (NREL) is used (see Dobos, 2014).

Finally, to compare the average power generations using the three DHI series, we present their boxplots in Figure 11.

The figure shows that the interquartile range of the calculated power using our proposed model is a superset of that using the observed data; however, this is not the case for the stationary model. Thus, our model can capture the uncertainties in power generations better than the stationary model. Furthermore, the median (mean) of the calculated power outputs obtained using our model is significantly closer to the true median (mean) than that obtained using the other model, implying that our model produces a more reliable power estimate.

Since the solar irradiance forecasts and simulated samples are directly employed by the power system operators, a highly reliable predictor and simulator will be beneficial; see Navid and Rosenwald (2013) and Li and Zhang (2020).

9 | DISCUSSION

In this study, we first introduced the solar DHI dataset of Saudi Arabia and established a broader perspective regarding the general nature of solar irradiances. Then, we showed that a time series of DHIs mostly possesses an underlying cyclostationary structure irrespective of the considered time range and location.

Next, we proposed a space-time model with temporal cyclostationarity and provided elaborate reasons for its construction. Thereafter, we discussed a simple but efficient two-step inference scheme and implemented it. Finally, we compared the proposed model with a stationary model in terms of predicting, simulating, and calculating power generations at new stations. In all these scenarios, the proposed model outperformed the stationary model.

Saudi Arabia is a country of desert and plenty of sunlight, which makes it highly suitable for producing considerable solar energy. Therefore, we believe that a reliable predictor and simulator of solar DHIs at new stations in Saudi Arabia will benefit the relevant authorities. Moreover, as mentioned in the introduction, the proposed model can be used to model solar DHIs measured in any other part of the world with slight or no modification. This provision makes the proposed model more practical and viable. Especially, in parts of the world where the cyclostationary structure in the DHI data is stronger than that of Saudi Arabia, like Sydney and Denver (see Subsection 3.1), the proposed model is likely to yield even better results.

CONFLICT OF INTEREST STATEMENT

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from K.A.CARE. Restrictions apply to the availability of these data, which were used under license for this study. Data are available from http://rratlas.kacare.gov.sa/ with the permission of K.A.CARE.

ORCID

Soumya Das https://orcid.org/0000-0002-6690-3128
Marc G. Genton https://orcid.org/0000-0001-6467-2998

REFERENCES

Abdulah, S., Ltaief, H., Sun, Y., Genton, M. G., & Keyes, D. E. (2018). ExaGeoStat: A high performance unified software for geostatistics on manycore systems. *IEEE Transactions on Parallel and Distributed Systems*, *29*(12), 2771–2784.

Alparslan, N. C., Kayabaşı, A., & Ruşen, S. E. (2019). Estimation of global solar radiation by using ANN and ANFIS. In *2019 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pp. 1–6.

AlYahya, S., & Irfan, M. A. (2016). Analysis from the new solar radiation atlas for Saudi Arabia. *Solar Energy*, *130*, 116–127.

Basawa, I. V., & Lund, R. (2001). Large sample properties of parameter estimates for periodic arma models. *Journal of Time Series Analysis*, *22*(6), 651–663.

Beiter, P., Elchinger, M., & Tian, T. (2017). 2016 renewable energy data book: National Renewable Energy Lab.(NREL), Golden, CO (United States).

Bloomfield, P., Hurd, H. L., & Lund, R. B. (1994). Periodic correlation in stratospheric ozone data. *Journal of Time Series Analysis*, *15*(2), 127–150.

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control*: John Wiley & Sons.

Broszkiewicz-Suwaj, E. (2003). Methods for determining the presence of periodic correlation based on the bootstrap methodology: Wrocław University of Technology, Tech. Rep. Research Report HSC/03/2.

Broszkiewicz-Suwaj, E., Makagon, A., Weron, R., & Wyłomańska, A. (2004). On detecting and modeling periodic correlation in financial data. *Physica A: Statistical Mechanics and its Applications*, *336*(1-2), 196–205.

Cressie, N. (1993). *Statistics for spatial data*: John Wiley & Sons.

Cressie, N., & Wikle, C. K. (2011). *Statistics for spatio-temporal data*: John Wiley & Sons.

Dandawate, A. V., & Giannakis, G. B. (1994). Statistical tests for presence of cyclostationarity. *IEEE Transactions on Signal Processing*, *42*(9), 2355–2369.

Das, S., Genton, M. G., Alshehri, Y. M., & Stenchikov, G. L. (2021). A cyclostationary model for temporal forecasting and simulation of solar global horizontal irradiance. *Environmetrics*, *32*, e2700.

Dawid, A. P. (1984). Present position and potential developments: Some personal views statistical theory the prequential approach. *Journal of the Royal Statistical Society: Series A (General)*, *147*(2), 278–290.

Dobos, A. P. (2014). Pvwatts version 5 manual: National Renewable Energy Lab. (NREL), Golden, CO (United States).

Filik, T., Filik, U. B., & Gerek, O. N. (2017). Solar radiation–to–power generation models for one-axis tracking PV system with on-site measurements from Eskisehir, Turkey. In *E3s web of conferences*, *22*, EDP Sciences, pp. 00046.

Genton, M. G. (2007). Separable approximations of space-time covariance matrices. *Environmetrics*, *18*(7), 681–695.

Hurd, H. L., & Gerr, N. L. (1991). Graphical methods for determining the presence of periodic correlation. *Journal of Time Series Analysis*, *12*(4), 337–350.

Hurd, H. L., & Miamee, A. (2007). *Periodically correlated random sequences: Spectral theory and practice*, Vol. 355: John Wiley & Sons.

Lauret, P., David, M., & Pedro, H. T. C. (2017). Probabilistic solar forecasting using quantile regression models. *Energies*, *10*(10), 1591.

Li, B., & Zhang, J. (2020). A review on the integration of probabilistic solar forecasting in power systems. *Solar Energy*, *207*, 777–795.

Lonij, V. P. A., Brooks, A. E., Cronin, A. D., Leuthold, M., & Koch, K. (2013). Intra-hour forecasts of solar power production using measurements from a network of irradiance sensors. *Solar Energy*, *97*, 58–66.

Mahmoudi, M. R., & Maleki, M. (2017). A new method to detect periodically correlated structure. *Computational Statistics*, *32*(4), 1569–1581.

Martin, D. E. K., & Kedem, B. (1993). Estimation of the period of periodically correlated sequences. *Journal of Time Series Analysis*, *14*(2), 193–205.

Myers, D. R., Wilcox, S., Anderberg, M., Al-Awaji, S. H., Al Abbadi, N. M., & bin Mahfoodh, M. Y. (1999). Saudi Arabian solar radiation network and data for validating satellite remote sensing systems, *3750*, Earth observing systems iv, pp. 503–513.

Napolitano, A. (2019). *Cyclostationary processes and time series: Theory, applications, and generalizations*: Academic Press.

Navid, N., & Rosenwald, G. (2013). Ramp capability product design for MISO markets. *Market Development and Analysis*. https://docplayer.net/56015209-Ramp-capability-product-design-for-miso-markets.html

Pagano, M. (1978). On periodic and multiple autoregressions. *The Annals of Statistics*, *6*(6), 1310–1317.

Pazheri, F. R. (2014). Solar power potential in Saudi Arabia. *International Journal of Engineering Research and Applications*, *4*(9), 171–174.

Pedro, H. T. C., & Coimbra, C. F. M. (2012). Assessment of forecasting techniques for solar power production with no exogenous inputs. *Solar Energy*, *86*(7), 2017–2028.

Pinson, P., & Kariniotakis, G. (2010). Conditional prediction intervals of wind power generation. *IEEE Transactions on Power Systems*, *25*(4), 1845–1856.

Pries, A., Ramírez, D., & Schreier, P. J. (2018). LMPIT-inspired tests for detecting a cyclostationary signal in noise with spatio-temporal structure. *IEEE Transactions on Wireless Communications*, *17*(9), 6321–6334.

Ranganai, E., & Sigauke, C. (2020). Capturing long-range dependence and harmonic phenomena in 24-hour solar irradiance forecasting: A quantile regression robustification via forecasts combination approach. *IEEE Access*, *8*, 172204–172218.

Reikard, G. (2009). Predicting solar radiation at high resolutions: A comparison of time series forecasts. *Solar Energy*, *83*(3), 342–349.

Salam, M. A., & Khan, S. A. (2018). Transition towards sustainable energy production–a review of the progress for solar energy in Saudi Arabia. *Energy Exploration & Exploitation*, *36*(1), 3–27.

Sun, Y., & Genton, M. G. (2011). Functional boxplots. *Journal of Computational and Graphical Statistics*, *20*(2), 316–334.

Taghezouit, B., Harrou, F., Sun, Y., Arab, A. H., & Larbes, C. (2021). A simple and effective detection strategy using double exponential scheme for photovoltaic systems monitoring. *Solar Energy*, *214*, 337–354.

Tsekouras, G., & Koutsoyiannis, D. (2014). Stochastic analysis and simulation of hydrometeorological processes associated with wind and solar energy. *Renewable Energy*, *63*, 624–633.

Van Loan, C. F. (2000). The ubiquitous Kronecker product. *Journal of Computational and Applied Mathematics*, *123*(1–2), 85–100.
Voyant, C., Notton, G., Kalogirou, S., Nivet, M.-L., Paoli, C., Motte, F., & Fouilloy, A. (2017). Machine learning methods for solar radiation forecasting: A review. *Renewable Energy*, *105*, 569–582.
Wang, J., Chen, T., & Huang, B. (2005). Cyclo-period estimation for discrete-time cyclostationary signals. *IEEE Transactions on Signal Processing*, *54*(1), 83–94.
Wikle, C. K., Zammit-Mangion, A., & Cressie, N. (2019). *Spatio-temporal statistics with r*: Chapman and Hall/CRC.
Wiser, R., & Bolinger, M. (2011). 2010 wind technologies market report: National Renewable Energy Lab.(NREL), Golden, CO (United States).
Yang, D., & van der Meer, D. (2021). Post-processing in solar forecasting: Ten overarching thinking tools. *Renewable and Sustainable Energy Reviews*, *140*, 110735.
Zell, E., Gasim, S., Wilcox, S., Katamoura, S., Stoffel, T., Shibli, H., Engel-Cox, J., & Al Subie, M. (2015). Assessment of solar radiation resources in Saudi Arabia. *Solar Energy*, *119*, 422–438.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Das, S., Alshehri, Y. M., Stenchikov, G. L., & Genton, M. G. (2023). A space-time model with temporal cyclostationarity for probabilistic forecasting and simulation of solar irradiance data. *Stat*, *12*(1), e583. https://doi.org/10.1002/sta4.583