

# A STOCHASTIC GENERATOR OF GLOBAL MONTHLY WIND ENERGY WITH TUKEY $g$ -AND- $h$ AUTOREGRESSIVE PROCESSES

Jaehong Jeong<sup>1</sup>, Yuan Yan<sup>2</sup>, Stefano Castruccio<sup>3</sup> and Marc G. Genton<sup>2</sup>

<sup>1</sup>*University of Maine*, <sup>2</sup>*King Abdullah University of Science and Technology*  
and <sup>3</sup>*University of Notre Dame*

*Abstract:* Quantifying the uncertainty of wind energy potential from climate models is a time-consuming task and requires considerable computational resources. A statistical model trained on a small set of runs can act as a stochastic approximation of the original climate model, and can assess the uncertainty considerably faster than by resorting to the original climate model for additional runs. While Gaussian models have been widely employed as means to approximate climate simulations, the Gaussianity assumption is not suitable for winds at policy-relevant (i.e., sub-annual) time scales. We propose a trans-Gaussian model for monthly wind speed that relies on an autoregressive structure with a Tukey  $g$ -and- $h$  transformation, a flexible new class that can separately model skewness and tail behavior. This temporal structure is integrated into a multi-step spectral framework that can account for global nonstationarities across land/ocean boundaries, as well as across mountain ranges. Inferences are achieved by balancing memory storage and distributed computation for a big data set of 220 million points. Once the statistical model was fitted using as few as five runs, it can generate surrogates rapidly and efficiently on a simple laptop. Furthermore, it provides uncertainty assessments very close to those obtained from all available climate simulations (40) on a monthly scale.

*Key words and phrases:* Big data, nonstationarity, spatio-temporal covariance model, sphere, stochastic generator, Tukey  $g$ -and- $h$  autoregressive model, wind energy.

## 1. Introduction

Wind energy is an important renewable energy source in many countries with no major negative environmental impacts (Wiser et al. (2011); Obama (2017)). Earth System Models (ESMs) provide physically consistent projections of wind energy potential, as well as spatially resolved maps in regions with poor observational coverage. However, these models are (more or less accurate) approximations of the actual state of the Earth's system, and the energy assessment

is therefore sensitive to changes in the model input. To address this, geoscientists generate a collection (*ensemble*) of ESMs to assess the sensitivity of the output (including wind) with respect to physical parameters and the trajectories of greenhouse gas concentrations (forcing scenarios). Recently, the role of the uncertainty due to ESMs' initial conditions (*internal variability*) has been identified as a prominent factor for multi-decadal projections, hence the importance of quantifying its uncertainty.

The Large ENSEMBLE (LENS) is a collection of 40 runs at the National Center for Atmospheric Research (NCAR) specifically designed to isolate the role of internal variability in the future climate (Kay et al. (2015)). The LENS required millions of CPU hours on a specialized supercomputer, and very few institutions have the resources and time for such an investigation. Is such an enormous task always necessary to assess internal variability? While it is absolutely necessary for quantities at the tail of the climate (e.g., temperature extremes), it is not always necessary for simpler indicators, such as the climate mean and variance. As part of a series of investigations promoted by KAUST on the topic of assessing wind energy in Saudi Arabia, Jeong et al. (2018) introduced the notion of a stochastic generator (SG), a statistical model that is trained on a small subset of LENS runs. The SG, or "Stochastic Generator of Climate Model Output"<sup>1</sup>, acts as a stochastic approximation of the climate model and, hence, allows for sampling more surrogate climate runs.<sup>2</sup> In their study, the authors present an SG for the global annual wind, and show that just five runs are sufficient to generate synthetic runs that are visually indistinguishable from the original simulations and have a similar spatio-temporal local dependence. However, while this SG introduced is able to approximate annual global data for the Arabian peninsula effectively, an annual scale is not useful for wind energy assessments. Thus, an SG at a finer temporal resolution for the same region is required to provide policy-relevant results.

An SG for monthly global wind output requires considerable modeling and computational effort. From a modeling perspective, data indexed on the sphere and over time require a dependence structure that can incorporate complex non-stationarities across the entire Earth system; see Jeong, Jun and Genton (2017) for a recent review of multiple approaches. For regularly spaced data, as is the case with atmospheric variables in an ESM output, multi-step spectrum models

---

<sup>1</sup>This is not to be confused with a Stochastic Weather Generator, which focuses on in-situ data at a high temporal resolution.

<sup>2</sup>A brief discussion on the difference between an SG and an emulator is contained in the same work.

are particularly useful. Such models can provide flexible nonstationary structures for Gaussian processes in the spectral domain, while maintaining the positive definiteness of the covariance functions (Jun and Stein (2008)). Recently, Castruccio and Guinness (2017) and Jeong et al. (2018) introduced a generalization that allows geographical descriptors, such as land/ocean indicators and mountain ranges, to be incorporated in a spatially varying spectrum.

In addition to the modeling complexity, the computational challenges are significant, because inferences need to be performed on a big data set. Over the last two decades, the increase in the size of spatio-temporal climate data sets has prompted the development of many new classes of scalable models. Of these, fixed rank methods (Cressie and Johannesson (2008)), predictive processes (Banerjee et al. (2008)), covariance tapering (Furrer, Genton and Nychka (2006)), and Gaussian Markov random fields (Rue and Held (2005)) have played a key role in our ability to couple the feasibility of an inference with the essential information to be communicated to stakeholders; see Sun, Li and Genton (2012) for a review. However, even by modern spatio-temporal data set standards, 220 million points is a considerable size. Thus, inferences require a methodology that leverages both parallel computing and the gridded geometry of the data. Castruccio and Genton (2018) provide a framework for a fast and parallel methodology for big climate data sets. However, the framework has thus far been limited to Gaussian processes. Whether an extension to non-Gaussian models with such a big data set is possible (and how) is an open question.

In this study, we propose an SG for monthly winds that is multi-step, spectral, and captures non-Gaussian behavior. We adopt a simple, yet flexible approach to construct non-Gaussian processes in time: the Tukey  $g$ -and- $h$  autoregressive process (Xu and Genton (2015); Yan and Genton (2019)), defined as  $Y(t) = \xi + \omega\tau_{g,h}\{Z(t)\}$ , where  $\xi$  is a location parameter,  $\omega$  is a scale parameter,  $Z(t)$  is a Gaussian autoregressive process, and  $\tau_{g,h}(z)$  is the Tukey  $g$ -and- $h$  transformation (Tukey (1977)):

$$\tau_{g,h}(z) = \begin{cases} g^{-1}\{\exp(gz) - 1\} \exp(hz^2/2) & \text{if } g \neq 0, \\ z \exp(hz^2/2) & \text{if } g = 0, \end{cases} \quad (1.1)$$

where  $g$  controls the skewness and  $h \geq 0$  governs the tail behavior. A significant advantage of Tukey  $g$ -and- $h$  autoregressive processes is that they provide flexible marginal distributions, allowing the skewness and heavy tails to be adjusted. This class of non-Gaussian processes is integrated within the multi-step spectral scheme to still allow inferences for a very big data set.

The remainder of the paper is organized as follows. Section 2 describes the wind data set. Section 3 details the statistical framework with the Tukey  $g$ -and- $h$  autoregressive models and the inferential approach. Section 4 provides a model comparison, and Section 5 illustrates how to generate SG runs. Section 6 concludes the paper.

## 2. The Community Earth System Model (CESM) Large ENSEMBLE (LENS) Project

We work on global wind data from LENS, which is an ensemble of CESM runs using version 5.2 of the Community Atmosphere Model of the NCAR (Kay et al. (2015)). The ensemble comprises runs at  $0.9375^\circ \times 1.25^\circ$  (latitude  $\times$  longitude) resolution, with each run under the Representative Concentration Pathway (RCP) 8.5 (van Vuuren et al. (2011)). Although the full ensemble consists of 40 runs, in our training set, we consider only  $R = 5$  randomly chosen runs for the SG to demonstrate that only a small number of runs is necessary (a full sensitivity analysis for  $R$  was performed in Jeong et al. (2018)). We consider the monthly near-surface wind speed at 10 m above ground level (U10 variable) for the period 2006 to 2100. Because our focus is on future wind trends, we analyze the projections for a total of 95 years. We consider all 288 longitudes, and discard latitudes near the poles to avoid numerical instabilities, consistent with previous works. These instabilities arise because of the close physical distance of neighboring points and the very different statistical behavior of wind speeds in the Arctic and Antarctic regions (McInnes, Erwin and Bathols (2011)). Therefore, we use 134 bands between  $62^\circ\text{S}$  and  $62^\circ\text{N}$ , and the full data set comprises approximately 220 million points ( $5 \times 1,140 \times 134 \times 288$ ). An example is given in Figure 1(a–d), where we show the ensemble mean and standard deviation of the monthly wind speed from the five selected runs, in March and September 2020. We observe that both means and standard deviations show temporal patterns. In particular, between the Tropic of Cancer and latitude  $60^\circ\text{N}$ , the mean wind speed over the ocean in September is stronger than that in March.

For each site, we test the significance of the skewness and kurtosis of the wind speed over time (Bai and Ng (2005)) after removing the climatology. In many spatial locations, the p-values are smaller than 0.05, as shown in Figure 1(e) and (f), indicating that the first two moments are not sufficient to characterize the temporal behavior of monthly wind over time. Most land points have significant skewness and, consistent with Bauer (1996), we observe that monthly wind

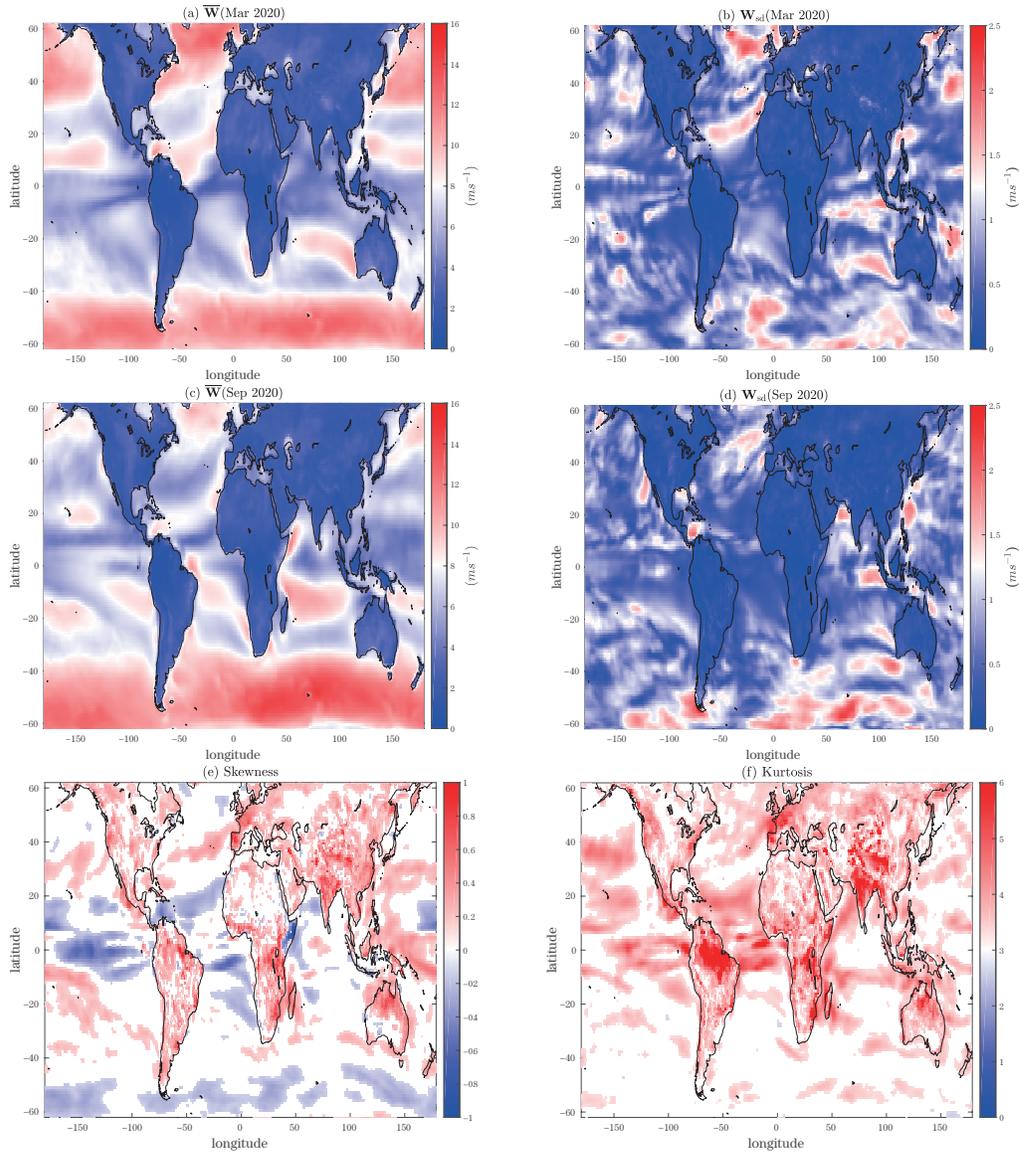


Figure 1. The (a) ensemble mean  $\overline{\mathbf{W}}(\text{March } 2020) = \sum_{r=1}^R \mathbf{W}_r(\text{March } 2020)/R$ , where  $R = 5$ , is the number of ensemble members, and (b) ensemble standard deviation  $\mathbf{W}_{\text{sd}}(\text{March } 2020) = \sqrt{\sum_{r=1}^R \{\mathbf{W}_r(\text{March } 2020) - \overline{\mathbf{W}}(\text{March } 2020)\}^2/R}$  of the monthly wind speed (in  $ms^{-1}$ ). (c) and (d) are the same as (a) and (b), but those in September 2020. The empirical skewness and kurtosis of the wind speed from one ensemble member after removing the trend are reported in (e) and (f), respectively, but only for locations where p-values of a significance test are less than 0.05.

speeds over the ocean are negatively skewed in the tropics, and positively skewed otherwise. The Tropical Indian Ocean and the Western Pacific Ocean, both areas with small wind speeds, are exceptions, with a positively skewed distribution.

### 3. The Space–Time Model

#### 3.1. The statistical framework

It is known that, after the climate model forgets its initial state, each ensemble member evolves in “deterministically chaotic” patterns (Lorenz (1963)). Climate variables in the atmospheric module have a tendency to forget their initial conditions after a short period, after which they evolve randomly, while still being attracted by the mean climate. Because ensemble members from the LENS differ only in their initial conditions (Kay et al. (2015)), we treat each one as a statistical realization from a common distribution. We define  $W_r(L_m, \ell_n, t_k)$  as the spatio-temporal monthly wind speed for realization  $r$  at the latitude  $L_m$ , longitude  $\ell_n$ , and time  $t_k$ , where  $r = 1, \dots, R$ ,  $m = 1, \dots, M$ ,  $n = 1, \dots, N$ , and  $k = 1, \dots, K$ , and define  $\mathbf{W}_r = \{W_r(L_1, \ell_1, t_1), \dots, W_r(L_M, \ell_N, t_K)\}^\top$ .

To remove the trend in our model, we consider  $\mathbf{D}_r = \mathbf{W}_r - \overline{\mathbf{W}}$ , with  $\overline{\mathbf{W}} = (1/R) \sum_{r=1}^R \mathbf{W}_r$ . The Gaussian assumption for  $\mathbf{D}_r$  is not, in general, valid at a monthly resolution (see Figure 1(e–f), Figure S1 for a significance test on the skewness and kurtosis, and Figure S2 for the Lilliefors and Jarque–Bera normality tests). Therefore, we apply the Tukey  $g$ -and- $h$  transformation (1.1), in which case, our model can be written as:

$$\mathbf{D}_r = \underline{\boldsymbol{\xi}} + \underline{\boldsymbol{\omega}} \cdot \boldsymbol{\tau}_{\mathbf{g}, \mathbf{h}}(\boldsymbol{\epsilon}_r), \quad \boldsymbol{\epsilon}_r \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}_{\text{space-time}})), \quad (3.1)$$

where  $\underline{\boldsymbol{\xi}} = \boldsymbol{\xi} \otimes \mathbf{1}_K$ , with  $\boldsymbol{\xi} = \{\xi(L_1, \ell_1), \dots, \xi(L_M, \ell_N)\}^\top$  a vector of the location parameters,  $\otimes$  the Kroneker product, and  $\mathbf{1}_K$  a vector of ones of length  $K$ . In addition,  $\underline{\boldsymbol{\omega}} = \boldsymbol{\omega} \otimes \mathbf{1}_K$ , with  $\boldsymbol{\omega} = \{\omega(L_1, \ell_1), \dots, \omega(L_M, \ell_N)\}^\top$  the vector of scale parameters,  $\underline{\mathbf{g}} = \mathbf{g} \otimes \mathbf{1}_K$ , and  $\underline{\mathbf{h}} = \mathbf{h} \otimes \mathbf{1}_K$ . Here,  $\mathbf{g} = \{g(L_1, \ell_1), \dots, g(L_M, \ell_N)\}^\top$  and  $\mathbf{h} = \{h(L_1, \ell_1), \dots, h(L_M, \ell_N)\}^\top$  the vectors of the  $MN$  parameters for the Tukey  $g$ -and- $h$  transformation at each site. Then,  $\boldsymbol{\tau}_{\mathbf{g}, \mathbf{h}}()$  represents the element-wise transformation according to (1.1). Therefore, component-wise, this becomes  $D_r(L_m, \ell_n, t_k) = \xi(L_m, \ell_n) + \omega(L_m, \ell_n) \tau_{g(L_m, \ell_n), h(L_m, \ell_n)}(\epsilon(L_m, \ell_n, t_k))$ .

We define  $\boldsymbol{\theta}_{\text{Tukey}} = \{\boldsymbol{\theta}_{T; m, n}\}_{m, n}$ , where  $m = 1, \dots, M$ ,  $n = 1, \dots, N$  and  $\boldsymbol{\theta}_{T; m, n} = \{\xi(L_m, \ell_n), \omega(L_m, \ell_n), g(L_m, \ell_n), h(L_m, \ell_n)\}^\top$  are the parameters of the Tukey  $g$ -and- $h$  transformation. Then, we define  $\boldsymbol{\theta}_{\text{space-time}} = (\boldsymbol{\theta}_{\text{time}}^\top, \boldsymbol{\theta}_{\text{lon}}^\top, \boldsymbol{\theta}_{\text{lat}}^\top)^\top$  as the vector of covariance parameters, which can be divided into temporal, longitudinal, and latitudinal dependence. The total set of parameters is  $\boldsymbol{\theta} =$

$(\boldsymbol{\theta}_{\text{Tukey}}^\top, \boldsymbol{\theta}_{\text{space-time}}^\top)^\top$ . Here,  $\boldsymbol{\theta}$  is very high dimensional. Hence, we consider a multi-step inference scheme, as first introduced by Castruccio and Stein (2013), where the parameters obtained from previous steps are assumed to be fixed and known:

Step 1. We estimate  $\boldsymbol{\theta}_{\text{Tukey}}$  and  $\boldsymbol{\theta}_{\text{time}}$  by assuming that there is no cross-temporal dependence in latitude and longitude.

Step 2. We consider  $\boldsymbol{\theta}_{\text{Tukey}}$  and  $\boldsymbol{\theta}_{\text{time}}$  fixed at their estimated values and estimate  $\boldsymbol{\theta}_{\text{lon}}$  by assuming that the latitudinal bands are independent.

Step 3. We consider  $\boldsymbol{\theta}_{\text{Tukey}}$ ,  $\boldsymbol{\theta}_{\text{time}}$ , and  $\boldsymbol{\theta}_{\text{lon}}$  fixed at their estimated values and estimate  $\boldsymbol{\theta}_{\text{lat}}$ .

This conditional step-wise approach implies some degree of error and uncertainty propagation across stages. Castruccio and Guinness (2017) provide guidelines on how to control for the propagation by using intermediate steps within Step 3. Following the same scheme, we detail the model for each of the three steps and also describe the inference.

**3.2. Step 1: Temporal dependence and inference for the Tukey  $g$ -and- $h$  model**

We assume that  $\boldsymbol{\epsilon}_r = \{\boldsymbol{\epsilon}_r(t_1)^\top, \dots, \boldsymbol{\epsilon}_r(t_K)^\top\}^\top$  in (3.1) evolves according to a vector autoregressive model of order  $p$  (VAR( $p$ )), with different parameters for each spatial location:

$$\begin{aligned} \boldsymbol{\epsilon}_r(t_k) &= \boldsymbol{\Phi}_1 \boldsymbol{\epsilon}_r(t_{k-1}) + \dots + \boldsymbol{\Phi}_p \boldsymbol{\epsilon}_r(t_{k-p}) + \mathbf{S} \mathbf{H}_r(t_k), \\ \mathbf{H}_r(t_k) &\stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{C}(\boldsymbol{\theta}_{\text{lon}}, \boldsymbol{\theta}_{\text{lat}})), \end{aligned} \tag{3.2}$$

where  $\boldsymbol{\Phi}_1 = \text{diag}\{\phi_{L_m, \ell_n}^1\}, \dots, \boldsymbol{\Phi}_p = \text{diag}\{\phi_{L_m, \ell_n}^p\}$  are  $MN \times MN$  diagonal matrices with autoregressive coefficients, and  $\mathbf{S} = \text{diag}\{S_{L_m, \ell_n}\}$  is a diagonal matrix of the standard deviations. Such a model assumes that there is no cross-temporal dependence across locations. Figure S3(a) in the Supplementary Material provides diagnostics on this assumption. Hence, at each spatial location, we have a Tukey  $g$ -and- $h$  autoregressive process of order  $p$  (Yan and Genton (2019)). The vector of temporal parameters is therefore  $\boldsymbol{\theta}_{\text{time}} = \{\boldsymbol{\theta}_{t;m,n}\}_{m,n}$ , where  $\boldsymbol{\theta}_{t;m,n} = (\phi_{L_m, \ell_n}^1, \dots, \phi_{L_m, \ell_n}^p, S_{L_m, \ell_n})^\top$ . The vectors  $\boldsymbol{\theta}_{\text{Tukey}}$  and  $\boldsymbol{\theta}_{\text{time}}$  in (3.1) are estimated simultaneously using a maximum approximated likelihood estimation (MALE, Xu and Genton (2015)). Furthermore, because the model assumes no cross-temporal dependence,  $\boldsymbol{\theta}_{T;m,n}$  and  $\boldsymbol{\theta}_{t;m,n}$  can be estimated independently and in parallel across  $m$  and  $n$ .

Exact likelihood inferences for the Tukey  $g$ -and- $h$  autoregressive process are computationally expensive because the inverse Tukey  $g$ -and- $h$  transformation  $\tau_{g,h}^{-1}$  does not have an explicit form (except when either  $g$  or  $h$  is equal to zero). The idea of the MALE is to approximate  $\tau_{g,h}^{-1}$  using a piecewise linear function  $\tilde{\tau}_{g,h}^{-1}$ , which reduces the computational time considerably compared with calculating  $\tau_{g,h}^{-1}$  numerically for each iteration in the optimization. The approximated log-likelihood function  $\tilde{l}$  of the monthly residual wind speed time series,  $\mathbf{D}_r(L_m, \ell_n) = \{D_r(L_m, \ell_n, t_1), \dots, D_r(L_m, \ell_n, t_K)\}^\top$ , from ensemble  $r$  at latitude  $L_m$  and longitude  $\ell_n$ , can be written as:

$$\begin{aligned} \tilde{l}(\boldsymbol{\theta}_{T;m,n}, \boldsymbol{\theta}_{t;m,n} \mid \mathbf{D}_r(L_m, \ell_n), ) \\ = f_{m,n}(\epsilon_1) + f_{m,n}(\epsilon_2 \mid \epsilon_1) + \dots + f_{m,n}(\epsilon_K \mid \epsilon_{K-1}, \dots, \epsilon_{K-p}) \\ - K \log\{\omega(L_m, \ell_n)\} - \frac{h(L_m, \ell_n)}{2} \sum_{k=1}^K \epsilon_k^2 \\ - \sum_{k=1}^K \log \left( \exp\{g(L_m, \ell_n)\epsilon_k\} + \frac{h(L_m, \ell_n)}{g(L_m, \ell_n)} [\exp\{g(L_m, \ell_n)\epsilon_k\} - 1] \epsilon_k \right), \end{aligned} \quad (3.3)$$

where  $\epsilon_k = \tilde{\tau}_{g(L_m, \ell_n), h(L_m, \ell_n)}^{-1} \{(D_r(L_m, \ell_n, t_k) - \xi(L_m, \ell_n)) / \omega(L_m, \ell_n)\}$  and  $f_{m,n}(\cdot) / f_{m,n}(\cdot \mid \cdot)$  is the corresponding marginal/conditional Gaussian log-likelihood for the underlying Gaussian AR( $p$ ) process with parameters  $\boldsymbol{\theta}_{t;m,n}$ .

For each location and each ensemble, the MALE is obtained by maximizing (3.3) with the optimal order  $p$  selected by the Bayesian information criterion (BIC). The results are shown in Figure S3(b). For a substantial share of points (56.1%),  $p > 0$  was selected, underscoring the need for a model with temporal dependence, even after differencing the original data from the average across realizations. A map of  $\hat{\phi}_{L_m, \ell_n}^1$ ,  $\hat{\phi}_{L_m, \ell_n}^2$ , and  $\hat{\phi}_{L_m, \ell_n}^3$  is shown in the Supplementary Material (Figure S4), along with the p-values (Figure S5).

Estimated values for  $\hat{\boldsymbol{\theta}}_{\text{Tukey}}$  are shown in Figure 2. Here,  $\hat{g}(L_m, \ell_n)$  and  $\hat{h}(L_m, \ell_n)$  were estimated with significant nonzero values over many locations (see Figure S6 for the p-value), and it is apparent how the Gaussian autoregressive model is not suitable for modeling monthly wind speed.

Once all parameters are estimated, the residuals can be calculated, as follows:

$$\begin{aligned} \hat{H}_r(L_m, \ell_n, t_k) = \frac{1}{\hat{S}(L_m, \ell_n)} \left\{ \hat{\epsilon}_r(L_m, \ell_n, t_k) - \hat{\phi}_{L_m, \ell_n}^1 \hat{\epsilon}_r(L_m, \ell_n, t_{k-1}) \right. \\ \left. - \dots - \hat{\phi}_{L_m, \ell_n}^p \hat{\epsilon}_r(L_m, \ell_n, t_{k-p}) \right\}, \end{aligned} \quad (3.4)$$

where  $\hat{\epsilon}_r(L_m, \ell_n, t_k) = \hat{\tau}_{\hat{g}(L_m, \ell_n), \hat{h}(L_m, \ell_n)}^{-1} [\{D_r(L_m, \ell_n, t_k) - \hat{\xi}(L_m, \ell_n)\} / \hat{\omega}(L_m, \ell_n)]$ ,

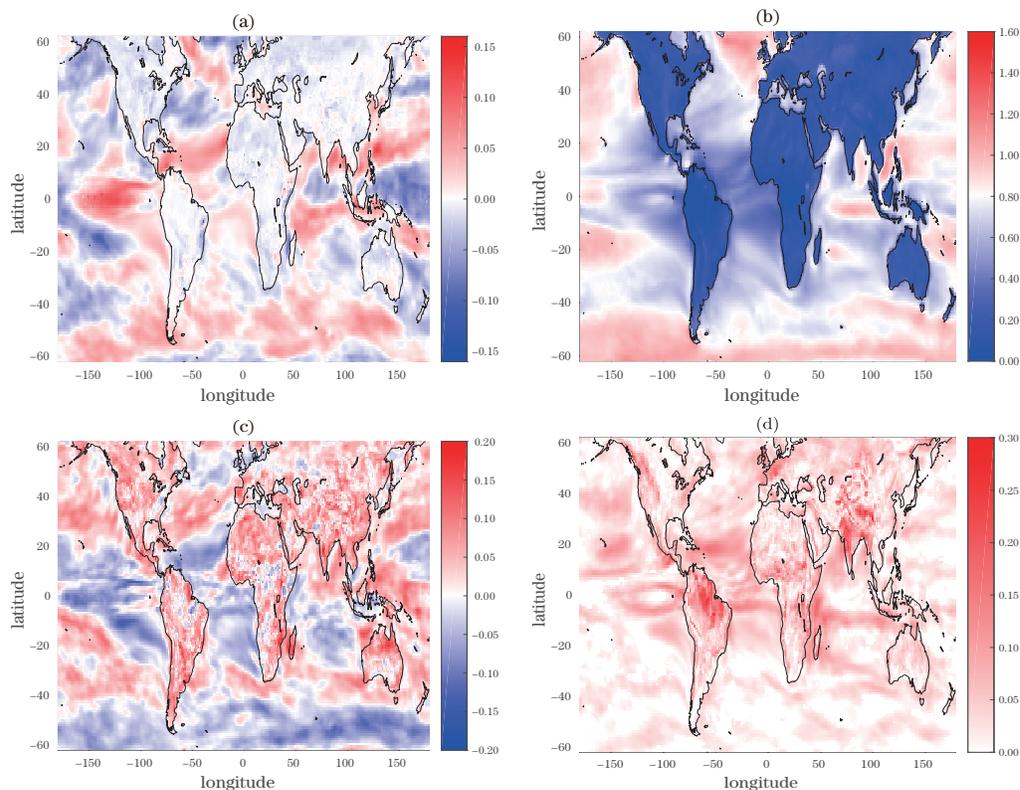


Figure 2. Plot of the estimated parameters  $\hat{\theta}_{\text{Tukey}}$  for the Tukey  $g$ -and- $h$  transformation, (a) location, (b) scale, (c)  $g$ , and (d)  $h$ .

and  $\hat{\tau}^{-1}_{\hat{g}(L_m, \ell_n), \hat{h}(L_m, \ell_n)}$  denotes the inverse Tukey  $g$ -and- $h$  transformation at latitude  $L_m$  and longitude  $\ell_n$ .

The following sections provide a model for the dependence structure of  $\mathbf{H}_r(t_k)$ , that is, a parametrization of  $\mathbf{C}(\theta_{\text{lon}}, \theta_{\text{lat}})$  in (3.2). Specifying a valid model for the entire spherical domain that captures global dependence structures is a nontrivial task. However, the following steps rely on the Gaussianity of  $\mathbf{H}_r(t_k)$ , and hence require only that we specify the covariance structure.

### 3.3. Step 2: Longitudinal structure

Here, we focus on  $\theta_{\text{lon}}$ ; that is, we provide a model for the dependence structure at different longitudes, but at the same latitude. Because the points are equally spaced and on a circle, the implied covariance matrix is circulant under a stationarity assumption (Davis (1979)), and is more naturally expressed in the spectral domain. The wind behavior on a latitudinal band, however, is

not longitudinally stationary. Recently, an evolutionary spectrum approach that allows for changing behavior across large-scale geographical descriptors was implemented successfully for global annual temperature and wind speed ensembles (Castruccio and Guinness (2017); Jeong et al. (2018); Castruccio and Genton (2018)). Here, we use a similar approach, and model  $H_r(L_m, \ell_n, t_k)$  in the spectral domain using a generalized Fourier transform across longitude. Indeed, if we define  $\iota = \sqrt{-1}$  to be the imaginary unit and  $c = 0, \dots, N - 1$  the wavenumber, then the process can be represented spectrally as

$$H_r(L_m, \ell_n, t_k) = \sum_{c=0}^{N-1} f_{L_m, \ell_n}(c) \exp(\iota \ell_n c) \tilde{H}_r(c, L_m, t_k), \quad (3.5)$$

where  $f_{L_m, \ell_n}(c)$  is a spectrum evolving across longitude, and  $\tilde{H}_r(c, L_m, t_k)$  is the spectral process.

To better account for the statistical behavior of wind speed, we implement a spatially varying model in which ocean, land, and high mountains above 1,000 m (following Jeong et al. (2018)) are treated as covariates. Therefore,  $f_{L_m, \ell_n}(c)$  depends on  $\ell_n$  being in a land, ocean, or high mountain domain, with the following expression:

$$f_{L_m, \ell_n}(c) = \begin{cases} f_{L_m, \ell_n}^1(c) & \text{if } (L_m, \ell_n) \in \text{high mountain,} \\ f_{L_m, \ell_n}^2(c) b_{\text{land}}(L_m, \ell_n; g'_{L_m}, r'_{L_m}) & \text{if } (L_m, \ell_n) \in \text{land,} \\ f_{L_m, \ell_n}^3(c) \{1 - b_{\text{land}}(L_m, \ell_n; g'_{L_m}, r'_{L_m})\} & \text{if } (L_m, \ell_n) \in \text{ocean,} \end{cases} \quad (3.6)$$

where  $b_{\text{land}}(L_m, \ell_n; g'_{L_m}, r'_{L_m}) = \sum_{n'=1}^N \tilde{I}_{\text{land}}(L_m, \ell_n; g'_{L_m}) w(L_m, \ell_n - \ell_{n'}; r'_{L_m})$  is a smooth function (taper) that allows a transition between land and ocean domains. Each of the three components of the spectrum in (3.6) is parametrized as follows (Castruccio and Stein (2013); Poppick and Stein (2014)):  $|f_{L_m, \ell_n}^j(c)|^2 = \psi_{L_m, \ell_n}^j \{(\alpha_{L_m, \ell_n}^j)^2 + 4 \sin^2(c\pi/N)\}^{-\nu_{L_m, \ell_n}^j}$ , for  $j = 1, 2, 3$ , where  $(\psi_{L_m, \ell_n}^j, \alpha_{L_m, \ell_n}^j, \nu_{L_m, \ell_n}^j)$  are interpreted as the variance, inverse range, and smoothness parameters, respectively, similarly to the Matérn spectrum. The parameters are modeled so that their logarithm changes continuously and depends linearly on the altitude; that is,  $\psi_{L_m, \ell_n}^j = \beta_{L_m}^{j, \psi} \exp[\tan^{-1}\{A_{L_m, \ell_n} \gamma_{L_m}^\psi\}]$ , for  $j = 1, 2$ , and  $\psi_{L_m, \ell_n}^3 = \beta_{L_m}^{3, \psi}$ , where  $\beta_{L_m}^{j, \psi} > 0$ ,  $\gamma_{L_m}^\psi \in \mathbb{R}$ , and  $A_{L_m, \ell_n}$  is the altitude at location  $(L_m, \ell_n)$ . Similar notation holds for  $\alpha_{L_m, \ell_n}^j$  and  $\nu_{L_m, \ell_n}^j$ . Hence, the longitudinal parameters are  $\boldsymbol{\theta}_{\text{lon}} = \{\boldsymbol{\theta}_{\ell, m}\}_m$ , where  $\boldsymbol{\theta}_{\ell, m} = \{(\beta_{L_m}^{j, \psi}, \gamma_{L_m}^\psi, \beta_{L_m}^{j, \alpha}, \gamma_{L_m}^\alpha, \beta_{L_m}^{j, \nu}, \gamma_{L_m}^\nu, g'_{L_m}, r'_{L_m})^\top, j = 1, 2, 3\}$ . Because the  $\boldsymbol{\theta}_{\ell, m}$  are independent across  $m$ , a model inference across latitudes can be performed independently using distributed computing.

To estimate the parameters for this and the next step, we leverage on the normality of the residuals  $H_r(L_m, \ell_n, t_k)$  and their independence across  $r$ . Indeed, by defining  $\mathbf{H} = (\mathbf{H}_1^\top, \dots, \mathbf{H}_R^\top)^\top$ , we can provide a restricted likelihood in closed form at latitude  $m$  (Castruccio and Stein (2013)):

$$2l(\boldsymbol{\theta}_{\ell,m} \mid \mathbf{H}) = KN(R - 1) \log(2\pi) + KN \log(R) \tag{3.7}$$

$$+ (R - 1)K \log |\boldsymbol{\Sigma}(\boldsymbol{\theta}_{\ell,m})| + \sum_{k=1}^K \sum_{r=1}^R \mathbf{H}_r^\top(t_k) \boldsymbol{\Sigma}(\boldsymbol{\theta}_{\ell,m}) \mathbf{H}_r(t_k),$$

where  $\boldsymbol{\Sigma}(\boldsymbol{\theta}_{\ell,m})$  is the  $N \times N$  covariance matrix of latitudinal band  $L_m$ , as implied by (3.5) and (3.6).

**3.4. Step 3: Latitudinal structure**

We now provide a model for the latitudinal dependence. Note that because the model in (3.5) is independent and identically distributed across  $r$  and  $t_k$ , we omit these two indices for simplicity. Castruccio and Stein (2013) and later works have proposed an autoregressive model for  $\tilde{H}(c, L_m)$  across  $m$  (but independent across  $c$ ). However, we consider a more general vector autoregressive model of order 1 (VAR(1)), such that  $\tilde{H}(c, L_m)$  is also allowed to depend on neighboring wavenumbers. We define  $\tilde{\mathbf{H}}_{L_m} = \{\tilde{H}(1, L_m), \dots, \tilde{H}(N, L_m)\}^\top$  and the latitudinal dependence by  $\tilde{\mathbf{H}}_{L_m} = \boldsymbol{\varphi}_{L_m} \tilde{\mathbf{H}}_{L_{m-1}} + \mathbf{e}_{L_m}$ , where  $\mathbf{e}_{L_m} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{L_m})$ , and  $\boldsymbol{\varphi}_{L_m}$  is an  $N \times N$  matrix containing the coefficients of the autoregressive structure across latitude;  $\boldsymbol{\Sigma}_{L_m}$  encodes the dependence for the innovation. To balance flexibility with computational feasibility, we seek a sufficiently sparse, but articulated structure for  $\boldsymbol{\varphi}_{L_m}$ . We propose a banded diagonally dominant matrix parametrized by  $a_{L_m}, b_{L_m} \in (-1, 1)$ , for all  $m$  values (the explicit expression is available in the Supplementary Material),  $\boldsymbol{\Sigma}_{L_m} = \text{diag}\{1 - \varphi_{L_m}(c)^2\}$  and  $\varphi_{L_m}(c) = \zeta_{L_m} \{1 + 4 \sin^2(c\pi/N)\}^{-\eta_{L_m}}$ , where  $\zeta_{L_m} \in [0, 1]$  and  $\eta_{L_m} > 0$  for all  $m$ . Hence, the latitudinal parameters are  $\boldsymbol{\theta}_{\text{lat}} = \{(a_{L_m}, b_{L_m}, \zeta_{L_m}, \eta_{L_m})^\top, m = 1, \dots, M\}$ .

We consider 10 sequential subsamples of 95 years (10 years each, except for the last partition) to reduce the computational burden. We derive an expression similar to (3.7) for this step, and estimate  $\zeta_{L_m}$  and  $\eta_{L_m}$  from each of the 10 subsamples, as shown in Figure S7 (other estimates of the longitudinal dependence parameters show similar patterns). Because there is no evidence of a change in latitudinal dependence over time, we consider the average of the parameter estimates. This value is used to combine multiple latitudinal bands and to generate surrogates in Section 5. The estimates  $\hat{a}_{L_m}$  and  $\hat{b}_{L_m}$  are also shown in Figure S8.

### 3.5. Computational aspects

An inference for a data set indexed by latitude, longitude, time, and realization comprising 220 million data points is a daunting task, even when using the aforementioned stepwise approach, which allows us to reduce the parameter space and to parallelize the likelihood maximization. Further mitigation of the computational and storage burden in step 3 is achievable by leveraging the gridded geometry of the data. Indeed, (3.7) can be expressed equivalently in the spectral domain using a fast Fourier transform of the data (Whittle likelihood, Whittle (1953)), such that the computational complexity is reduced from  $O(M^3N^3)$  to  $O(M^2N \log N)$  and the storage requirement is reduced from  $O(M^2N^2)$  to  $O(M^2N)$ .

We used a workstation with  $2 \times 12$  cores of Intel Xeon E5-2680V3 2.5GHz processors. Step 1 required approximately 6 hours, step 2 required 29 hours, and step 3 required 179 hours, for a total of approximately nine days. Inferences are therefore nontrivial and require considerable computational resources. However, once the parameters were estimated, generating the 40 statistical surrogates needed for Section 5 required only 16 minutes on a simple laptop (see the Matlab Graphical User Interface described in the application).

## 4. Model Comparison

To validate our proposed model based on the Tukey  $g$ -and- $h$  autoregressive (TGH-AR) process, we compare it with both a Gaussian autoregressive (G-AR) process and with two models with special cases of spatial dependence structure from steps 2 and 3, as detailed in Sections 3.3 and 3.4, respectively. In the Supplementary Material, we provide additional comparisons with a model with no spatial dependence and one with Gaussian dependence (Figures S9 and S10).

### 4.1. Comparison with a Gaussian temporal autoregressive process

In our first comparison, note that the G-AR process can be obtained from (3.1) by assuming  $\boldsymbol{\xi} = \mathbf{0}$ ,  $\boldsymbol{\omega} = \mathbf{1}$ ,  $\mathbf{g} = \mathbf{0}$ , and  $\mathbf{h} = \mathbf{0}$ ; therefore, a formal model selection can be performed. Figure 3 represents the BIC between the two models at each site from one ensemble member. Positive and negative values indicate a better and worse model fit of the TGH-AR compared with the G-AR, respectively. The TGH-AR outperforms the G-AR in more than 85% of the spatial locations, with a considerable improvement in the BIC score (the map scale is in the order of  $10^3$ ). Overall, the fit for land sites is considerably better for the TGH-AR, with

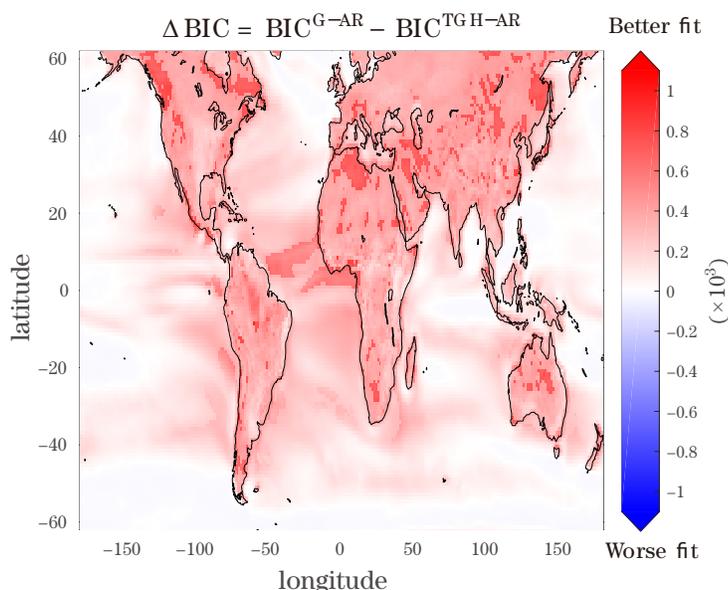


Figure 3. Map of differences in the BIC between the TGH-AR and G-AR from one ensemble member.

peaks in the North Africa area near Tunisia and in and around Saudi Arabia, in the region of study in Section 5. The tropical Atlantic also shows large gains.

#### 4.2. Comparison with submodels of global dependence

The TGH-AR model is also compared with a model with no altitude dependence, that is, where

$$\psi_{L_m, \ell_n}^1 = \psi_{L_m, \ell_n}^2, \alpha_{L_m, \ell_n}^1 = \alpha_{L_m, \ell_n}^2, \nu_{L_m, \ell_n}^1 = \nu_{L_m, \ell_n}^2 \implies f_{L_m, \ell_n}^1(c) = f_{L_m, \ell_n}^2(c),$$

for all  $m, n, c$  in (3.6). The model still assumes an evolutionary spectrum with changing behavior across land/ocean (Castruccio and Guinness (2017)), and is denoted by LAO. We further compare the TGH-AR with a model with an autoregressive dependence across latitudes, that is, a model in which  $a_{L_m} = b_{L_m} = 0$  in the parametrization of  $\varphi_{L_m}$  in Section 3.4, which we denote as ARL.

Because the LAO and ARL are both special cases of the TGH-AR, a formal comparison of their model selection metrics can be performed (see Table 1). There is evidence of a considerable improvement from the LAO to the ARL, indicating the need to incorporate the altitude when modeling the covariance structure. The additional smaller (although non-negligible, because the BIC improvement is approximately  $10^5$ ) improvement from the ARL to the TGH-

Table 1. Comparison of the number of parameters (excluding the temporal component), the normalized restricted log-likelihood, and BIC for three different models: LAO, ARL, and TGH-AR. The general guidelines for  $\Delta\loglik/\{NMK(R-1)\}$  are that values above 0.1 are considered to be large, and those above 0.01 are modest, but still sizable (Castuccio and Stein (2013)).

Model	LAO	ARL	TGH-AR
# of parameters	1,338	2,142	2,408
$\Delta\loglik/\{NMK(R-1)\}$	0	0.0440	<b>0.0443</b>
BIC ( $\times 10^8$ )	-5.8963	-6.0511	<b>-6.0521</b>

Table 2. 25th, 50th, and 75th percentiles of two difference metrics over ocean, land, and high mountains near the Indian ocean.

Metric	Region	25th	50th	75th
$[\{\Delta_{ew;m,n} - \hat{\Delta}_{ew;m,n}^{\text{ARL}}\}^2 - \{\Delta_{ew;m,n} - \hat{\Delta}_{ew;m,n}^{\text{TGH-AR}}\}^2] \times 10^4$	ocean	0	0	0
	land	-14	0	16
	mountain	-8	5	22
$[\{\Delta_{ns;m,n} - \hat{\Delta}_{ns;m,n}^{\text{ARL}}\}^2 - \{\Delta_{ns;m,n} - \hat{\Delta}_{ns;m,n}^{\text{TGH-AR}}\}^2] \times 10^4$	ocean	-1	1	2
	land	-2	2	11
	mountain	-2	1	7

AR underscores the necessity of a flexible model that is able to account for dependence across both wavenumbers and latitudes.

All three models can also be compared using local contrasts, because the residuals in (3.4) are approximately Gaussian. We focus on the contrast variances to assess the goodness of fit of the model in terms of its ability to reproduce the local dependence (Jun and Stein (2008)):

$$\begin{aligned}\Delta_{ew;m,n} &= \frac{1}{KR} \sum_{k=1}^K \sum_{r=1}^R \{H_r(L_m, \ell_n, t_k) - H_r(L_m, \ell_{n-1}, t_k)\}^2, \\ \Delta_{ns;m,n} &= \frac{1}{KR} \sum_{k=1}^K \sum_{r=1}^R \{H_r(L_m, \ell_n, t_k) - H_r(L_{m-1}, \ell_n, t_k)\}^2,\end{aligned}\tag{4.1}$$

where  $\Delta_{ew;m,n}$  and  $\Delta_{ns;m,n}$  denote the east–west and north–south contrast variances, respectively.

We compare the ARL with the TGH-AR, and compute the squared distances between the empirical and fitted variances. We find that the TGH-AR shows a better model fit in the case of the north–south contrast variance, but that there is no noticeable difference between the two models in the case of the east–west variance. A representation of these differences for the small region of interest near South Africa ( $13.75^\circ\text{E} \sim 48.75^\circ\text{E}$  and  $30^\circ\text{S} \sim 4^\circ\text{N}$ ) is given in Figure S11.

Positive values are obtained when the TGH-AR is a better model fit than is the ARL; negative values are obtained when the ARL is the better model fit. Figure S11(a) and (b) shows that dark red colors are more widely spread over mountains and that no clear difference is shown over the ocean. The results presented in Table 2 are consistent with the visual inspection. In addition, the two metrics, particularly, over mountainous areas, show larger values than those obtained for the ocean areas. In a global mean or median of the metrics, there is no significant difference between the two models.

## 5. Generation of Stochastic Surrogates

Once the model is properly defined and validated, we apply it to produce surrogate runs and train the SG with  $R = 5$  climate runs. A comprehensive sensitivity analysis on the number of elements in the training set can be found in Jeong et al. (2018). We use the SG to assess the uncertainty of the monthly wind power density (WPD), and compare it with the results of the full extent of the LENS runs.

The mean structure of the model is obtained by smoothing the ensemble mean  $\overline{\mathbf{W}}$ , but such an estimate is highly variable. For each latitude and longitude (i.e., each  $n$  and  $m$ ), we fit a spline  $\widetilde{W}(L_m, \ell_n, t_k)$  that minimizes the following function (Castruccio and Guinness (2017); Jeong et al. (2018)):  $\lambda \sum_{k=1}^K \left\{ \overline{W}(L_m, \ell_n, t_k) - \widetilde{W}(L_m, \ell_n, t_k) \right\}^2 + (1 - \lambda) \sum_{k=1}^K \left\{ \nabla_2 \widetilde{W}(L_m, \ell_n, t_k) \right\}^2$ , where  $\nabla_2$  is the discrete Laplacian. We impose  $\lambda = 0.99$  to give significant weight to the spline interpolant in order to reflect the varying patterns of monthly wind fields over the next century. For each spatial location, a harmonic regression of a time series may also be used to estimate the mean structure, but for the sake of simplicity, we opt for a nonparametric description. Once  $\boldsymbol{\theta} = (\boldsymbol{\theta}_{\text{Tukey}}^\top, \boldsymbol{\theta}_{\text{space-time}}^\top)^\top$  is estimated from the training set, surrogate runs can be generated easily using Algorithm 1.

We generate 40 SG runs using the proposed model and compare them with the original 40 LENS runs. As clearly shown in Figures 1(a) and S12(a), the ensemble means from the training set and the SG runs are visually indistinguishable.

We also evaluate both models in terms of their structural similarity index; to that end, we compare local patterns of pixel intensities that have been standardized for luminance and contrast (Figure S13) (Wang et al. (2004); Castruccio, Genton and Sun (2019)). We observe that the SG runs from the Tukey  $g$ -and-

**Algorithm 1** Generate surrogates

- 
- 1: **procedure** GENERATE SURROGATES
  - 2: Generate  $\mathbf{e}_{L_m} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{L_m})$  as in Section 3.4.
  - 3: Compute the VAR(1) process  $\tilde{\mathbf{H}}_{L_m}$  as in Section 3.4.
  - 4: Compute  $H_r(L_m, \ell_n, t_k)$  from (3.5)
  - 5: Compute  $\boldsymbol{\epsilon}_r$  with equation (3.2), and obtain  $\tilde{\mathbf{D}}_r$  from the Tukey  $g$ -and- $h$  transformation (3.1)
  - 6: Obtain the SG run as  $\tilde{\mathbf{W}} + \tilde{\mathbf{D}}_r$ , where  
 $\tilde{\mathbf{W}} = \{\tilde{W}(L_1, \ell_1, t_1), \dots, \tilde{W}(L_M, \ell_1, t_1), \tilde{W}(L_1, \ell_2, t_1), \dots, \tilde{W}(L_M, \ell_N, t_K)\}^\top$ .
  - 7: **end procedure**
- 

$h$  case produce maps that are visually more similar to the original LENS runs than those in the Gaussian case (see also Figure S14 and S15 for the measures of skewness and kurtosis, and Figure S16 for a visual comparison of the runs in one location).

We further compare the LENS and SG in terms of the near-future trend (2013–2046), a reference metric for the LENS (Kay et al. (2015)) that was used to illustrate the influence of the internal variability on global warming trends. We compute the near-future wind speed trends near the Indian ocean for each of the SG and LENS runs. The results are shown in Figure 4(a) and (b). One can clearly see that the mean near-future wind trends by the SG runs are very similar to those from the training set of LENS runs.

Next, we assess the wind energy potential. The WPD (in  $Wm^{-2}$ ) evaluates the wind energy resource available at the site for conversion by a wind turbine. The WPD can be calculated as  $WPD = 0.5\rho u^3$ ,  $u = u_r(z/z_r)^\alpha$ , where  $\rho$  is the air density ( $\rho = 1.225 \text{ kgm}^{-3}$  in this study),  $u$  is the wind speed at a certain height  $z$ ,  $u_r$  is the known wind speed at a reference height  $z_r$ , and  $\alpha = 1/7$  (Peterson and Hennessey Jr (1978); Newman and Klein (2013)). We focus our analysis on the Gulf of Aden ( $46.25^\circ\text{E}$  and  $12.72^\circ\text{N}$ ), a narrow channel connecting the Red Sea to the Indian Ocean characterized by high wind regimes (Yip, Gunturu and Stenchikov (2017)). In addition, we choose to work on the WPD in 2020 at 80 m, a standard height for wind turbines (Holt and Wang (2012); Yip, Gunturu and Stenchikov (2017)).

For completeness, we also considered Gaussian-based SG runs. We refer to these as the SG-G runs, and our original SG runs as the SG-T runs to distinguish between the two. The results for March and September 2020 are shown in Figure 4(c,d) and (e,f), with the histograms representing both the SG-G and

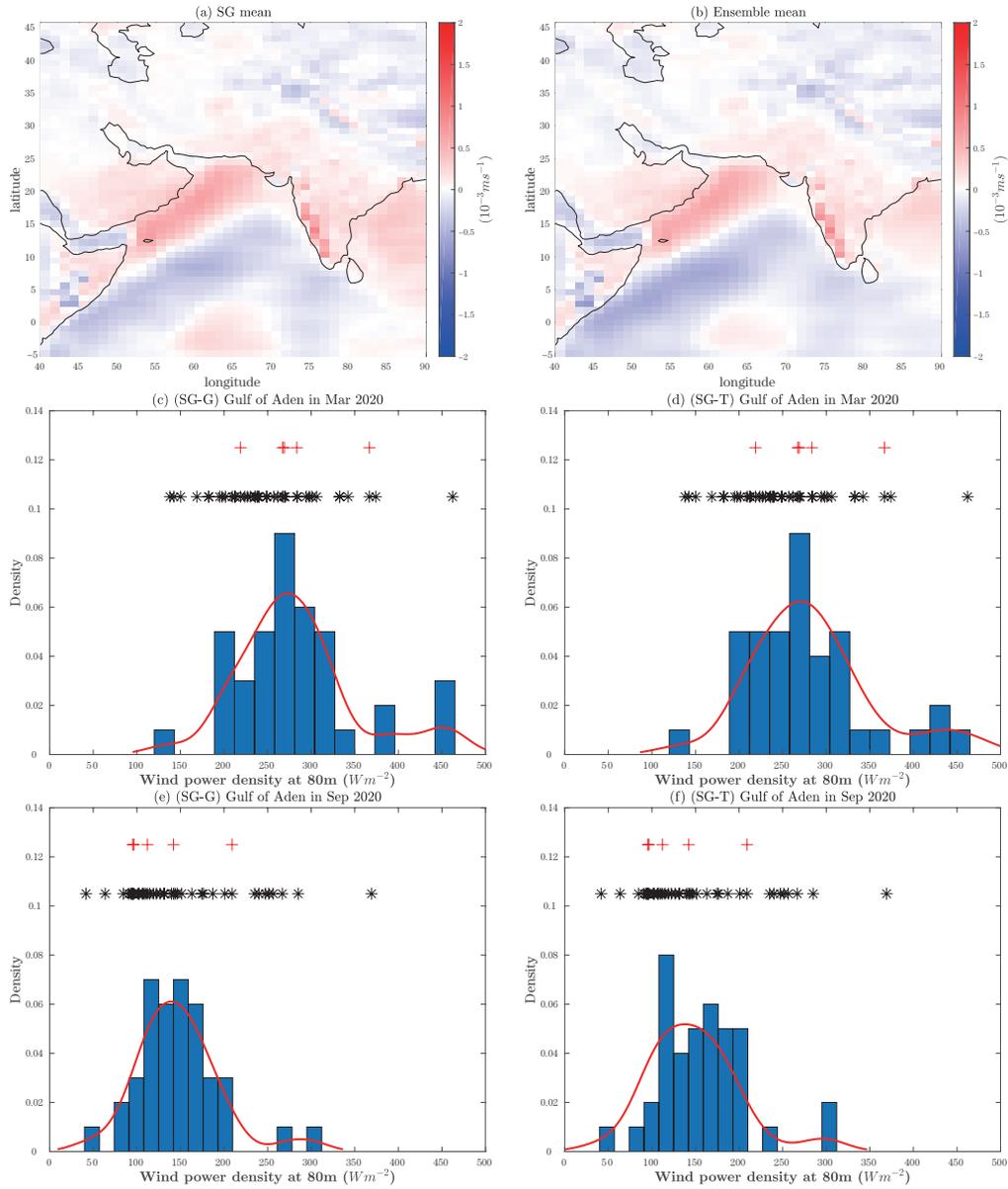


Figure 4. Maps of (a) the mean from the SG runs and (b) the ensemble mean from the near-future (2013–2046) near-surface wind speed trends near the Indian ocean. Histogram of the distribution of the WPD at 80 m, with the nonparametric density in red for the 40 SG-G and SG-T runs near the Gulf of Aden (c,d) in March 2020, and (e,f) in September 2020 (\* represents the LENS runs, + represents the five LENS runs in the training set of the SG).

the SG-T runs, a superimposed estimated nonparametric density in red, and the LENS runs on top with an asterisk marker. For both cases, all histograms have right-skewed distributions, as in the distribution of the entire LENS. It is clear that the distribution resulting from the SG runs is more informative than the five LENS runs in the training set (see red cross markers on top). Furthermore, it matches the uncertainty generated by the 40 LENS runs. Figure S17 shows a comparison for the same location and months in terms of QQ-plots for both our SG-T model and a model with no spatial dependence. It is apparent how the spatially dependent model results in a univariate fit closer to the LENS data in both months.

In Figure 5, we show the boxplots of the distribution of the WPD in 2020 for the LENS against the two SG runs across all months. The point estimates and ranges of the WPD values from the LENS runs are well-matched by those from the SGs, with slight misfits in April and November. The importance of such results cannot be understated: both SG runs are able to capture the inter-annual WPD patterns and its internal variability in a region of critical importance for wind farming. The internal variability in the months of high wind activity, such as July, is such that the WPD can be classified from fair to very high, according to standard wind energy categories (Archer and Jacobson (2003)). Furthermore, the SGs can reproduce the same range with as few as five runs in the training set. Overall, both SG runs perform comparatively well, but we find that the empirical skewness and kurtosis values from the SG-T runs are more similar to those of the 40 LENS runs than are the values from the SG-G runs. We computed the differences of the skewness and kurtosis values between the SG runs and LENS runs for each month in 2020. Then, we took an average (or median) of the absolute values of the differences across months. As a result, we obtained that the average (or median) metrics in the skewness values for the SG-G and SG-T runs are 0.3572 (0.3576) and 0.3142 (0.2151), respectively. In addition, the metrics in the kurtosis values were 0.8926 (0.5586) and 0.7948 (0.4531), respectively.

The generation of surrogate runs is fast and can be performed on a simple laptop, as long as the estimated parameters are provided. We have developed a Matlab Graphical User Interface (GUI, see Figure S18) that allows an end user to interactively generate and store several surrogate runs on a simple laptop in several minutes. The GUI is simple and intuitive, and requires only the stored estimated parameters, along with the algorithm described in this section for data generation. Thus, approximately 123 MB is required to generate as many

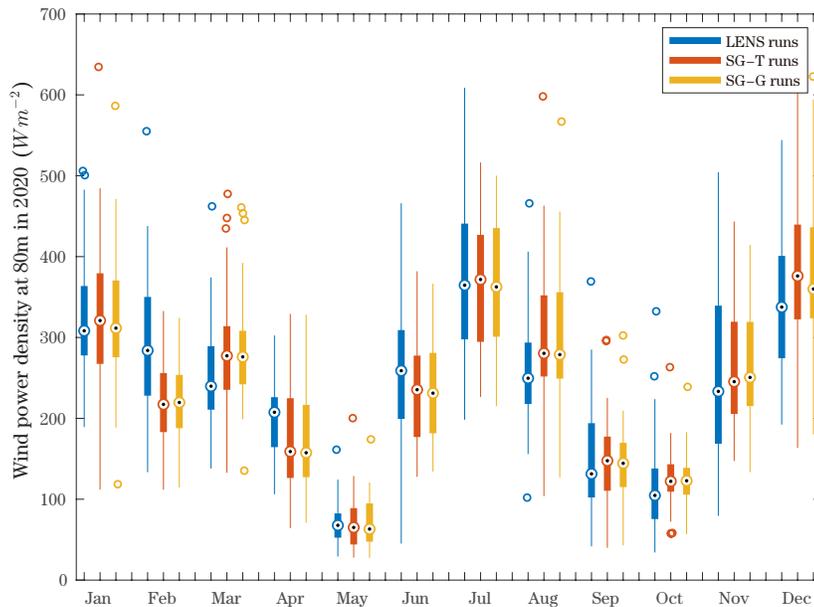


Figure 5. Boxplots of the distribution of the WPD at 80 m, in 2020, for 40 LENS runs and the 40 SG runs based on the Tukey  $g$ -and- $h$  and Gaussian cases near the Gulf of Aden.

ensembles as desired. In contrast, the storage of the five LENS runs required 1.7 GB.

## 6. Discussion and Conclusion

In this work, we proposed a non-Gaussian, multi-step spectral model for a global space-time data set of more than 220 million points. Motivated by the need to approximate computer output with a faster surrogate, we provided a fast, parallelizable, and scalable methodology to perform inferences on a big data set and to assess the uncertainty of global monthly wind energy.

Our proposed model relies on a trans-Gaussian process, the Tukey  $g$ -and- $h$ , which allows us to control the skewness and tail behavior using two distinct parameters. This class of models is embedded in a multi-step approach to allow for inferences for a nonstationary global model, while also capturing site-specific temporal dependence. Our results show that it clearly outperforms currently available Gaussian models.

Our model has been applied as an SG, a new class of stochastic approximations that uses global models to more efficiently assess the internal variability

of wind energy resources in developing countries with poor observational data coverage. Our results suggest that the uncertainty produced by the SG with a training set of five runs is very similar to that from 40 LENS runs in regions of critical interest for wind farming. Therefore, our model can be used as an efficient surrogate to assess the variability of wind energy at the monthly level, a clear improvement from the annual results presented by Jeong et al. (2018), and an important step forward using SGs at policy-relevant time scales.

Although we focused on global wind energy assessments, the use of SGs goes beyond the scope of this application. Indeed, similar models can and have been proposed in the literature to explore the sensitivity of temperature (Castruccio and Genton (2016)). The stepwise approach proposed in Section 3 can also be applied to data sets not related to geoscience, as long as the data suggest different scales of spatio-temporal dependence. For example, Castruccio, Ombao and Genton (2018) applied a similar stepwise approach to fMRI data, which showed spatial dependence at the voxel, regional, and whole-brain level.

## Supplementary Materials

The online Supplementary Material provides the additional results described in the text for Sections 2, 3.2, 3.4, 4, and 5, as well as a Matlab GUI.

## Acknowledgment

This publication is based upon work supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No: OSR-2015-CRG4-2640.

## References

- Archer, C. L. and Jacobson, M. Z. (2003). Spatial and temporal distributions of US winds and wind power at 80 m derived from measurements. *Journal of Geophysical Research: Atmospheres* **108**, 4289.
- Bai, J. and Ng, S. (2005). Tests for skewness, kurtosis, and normality for time series data. *Journal of Business & Economic Statistics* **23**, 49–60.
- Banerjee, S., Gelfand, A. E., Finley, A. O. and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 825–848.
- Bauer, E. (1996). Characteristic frequency distributions of remotely sensed in situ and modelled wind speeds. *International Journal of Climatology* **16**, 1087–1102.
- Castruccio, S. and Genton, M. G. (2016). Compressing an ensemble with statistical models: an algorithm for global 3D spatio-temporal temperature. *Technometrics* **58**, 319–328.

- Castruccio, S. and Genton, M. G. (2018). Principles for inference on big spatio-temporal data from climate models. *Statistics & Probability Letters* **136**, 92–96.
- Castruccio, S., Genton, M. G. and Sun, Y. (2019). Visualising spatio-temporal models with virtual reality: from fully immersive environments to apps in stereoscopic view (with discussion). *Journal of the Royal Statistical Society - Series A (Statistics in Society)* **182**, 379–387.
- Castruccio, S. and Guinness, J. (2017). An evolutionary spectrum approach to incorporate large-scale geographical descriptors on global processes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **66**, 329–344.
- Castruccio, S., Ombao, H. and Genton, M. G. (2018). A scalable multi-resolution spatio-temporal model for brain activation and connectivity in fMRI data. *Biometrics* **74**, 823–833.
- Castruccio, S. and Stein, M. L. (2013). Global space-time models for climate ensembles. *The Annals of Applied Statistics* **7**, 1593–1611.
- Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 209–226.
- Davis, P. J. (1979). *Circulant Matrices*. American Mathematical Society.
- Furrer, R., Genton, M. G. and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics* **15**, 502–523.
- Holt, E. and Wang, J. (2012). Trends in wind speed at wind turbine height of 80 m over the contiguous united states using the north american regional reanalysis (NARR). *Journal of Applied Meteorology and Climatology* **51**, 2188–2202.
- Jeong, J., Castruccio, S., Crippa, P. and Genton, M. G. (2018). Reducing storage of global wind ensembles with stochastic generators. *The Annals of Applied Statistics* **12**, 490–509.
- Jeong, J., Jun, M. and Genton, M. G. (2017). Spherical process models for global spatial statistics. *Statistical Science* **32**, 501–513.
- Jun, M. and Stein, M. L. (2008). Nonstationary covariance models for global data. *The Annals of Applied Statistics* **2**, 1271–1289.
- Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., Arblaster, J. M., Bates, S. C., Danabasoglu, G., Edwards, J., Holland, M., Kushner, P., Lamarque, J.-F., Lawrence, D., Lindsay, K., Middleton, A., Munoz, E., Neale, R., Oleson, K., Polvani, L. and Vertenstein, M. (2015). The community earth system model (cesm) large ensemble project: a community resource for studying climate change in the presence of internal climate variability. *Bulletin of the American Meteorological Society* **96**, 1333–1349.
- Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences* **20**, 130–141.
- McInnes, K. L., Erwin, T. A. and Bathols, J. M. (2011). Global climate model projected changes in 10 m wind speed and direction due to anthropogenic climate change. *Atmospheric Science Letters* **12**, 325–333.
- Newman, J. and Klein, P. (2013). Extrapolation of wind speed data for wind energy applications. In *Fourth Conference on Weather, Climate, and the New Energy Economy. Annual Meeting of the American Meteorological Society* 7. Austin, TX.
- Obama, B. (2017). The irreversible momentum of clean energy. *Science*.
- Peterson, E. W. and Hennessey Jr, J. P. (1978). On the use of power laws for estimates of wind power potential. *Journal of Applied Meteorology* **17**, 390–394.

- Poppick, A. and Stein, M. L. (2014). Using covariates to model dependence in nonstationary, high-frequency meteorological processes. *Environmetrics* **25**, 293–305.
- Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Boca Raton, FL: CRC Press.
- Sun, Y., Li, B. and Genton, M. G. (2012). Geostatistics for large datasets. In *Advances and Challenges in Space-Time Modelling of Natural Events* (Edited by E. Porcu, J. M. Montero and M. Schlather), 55–77. Springer.
- Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.
- van Vuuren, D. P., Edmonds, J., Kainuma, M., Riahi, K., Thomson, A., Hibbard, K., Hurtt, G. C., Kram, T., Krey, V., Lamarque, J.-F., Masui, T., Meinshausen, M., Nakicenovic, N., Smith, S. J. and Rose, S. K. (2011). The representative concentration pathways: an overview. *Climatic Change* **109**, 5–31.
- Wang, Z., Bovik, A. C., Sheikh, H. R. and Simoncelli, E. P. (2004). Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing* **13**, 600–612.
- Whittle, P. (1953). Estimation and information in stationary time series. *Arkiv för Matematik* **2**, 423–434.
- Wiser, R., Yang, Z., Hand, M., Hohmeyer, O., Infield, D., Jensen, P. H., Nikolaev, V., O'Malley, M., Sinden, G. and Zervos, A. (2011). Wind energy. In *IPCC Special Report on Renewable Energy Sources and Climate Change Mitigation*. (Edited by O. Edenhofer, R. Pichs-Madruga, Y. Sokona, K. Seyboth, P. Matschoss, S. Kadner, T. Zwickel, P. Eickemeier, G. Hansen and S.
- Xu, G. and Genton, M. G. (2015). Efficient maximum approximated likelihood inference for Tukey's  $g$ -and- $h$  Distribution. *Computational Statistics & Data Analysis* **91**, 78–91.
- Yan, Y. and Genton, M. G. (2019). Non-gaussian autoregressive processes with Tukey  $g$ -and- $h$  transformations. *Environmetrics* **30**, e2503.
- Yip, C. M. A., Gunturu, U. B. and Stenchikov, G. L. (2017). High-altitude wind resources in the middle east. *Scientific Reports* **7**, 9885.

Department of Mathematics and Statistics, University of Maine, Orono, ME 04469, USA.

E-mail: jaehong.jeong@maine.edu

Statistics Program, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia.

E-mail: yuan.yan@kaust.edu.sa

Department of Applied and Computational Mathematics and Statistics, 153 Hurley Hall, University of Notre Dame, Notre Dame, IN 46556, USA.

E-mail: scastruc@nd.edu

Statistics Program, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia.

E-mail: marc.genton@kaust.edu.sa

(Received November 2017; accepted October 2018)