Statistica Sinica Preprint No: SS-2017-0474	
Title	A Stochastic Generator of Global Monthly Wind Energy with Tukey g-and-h Autoregressive Processes
Manuscript ID	SS-2017-0474
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202017.0474
Complete List of Authors	Jaehong Jeong Yuan Yan Stefano Castruccio and Marc G. Genton
Corresponding Author	Marc G. Genton
E-mail	marc.genton@kaust.edu.sa
Notice: Accepted version subject to English editing.	

A Stochastic Generator of Global Monthly Wind Energy with Tukey g-and-h Autoregressive Processes

Jaehong Jeong[1], Yuan Yan[2], Stefano Castruccio[3], and Marc G. Genton[2]

October 5, 2018

Abstract

Quantifying the uncertainty of wind energy potential from climate models is a very time-consuming task and requires a considerable amount of computational resources. A statistical model trained on a small set of runs can act as a stochastic approximation of the original climate model, and be used to 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 time scales, i.e., sub-annual. We propose a trans-Gaussian model for monthly wind speed that relies on an autoregressive structure with 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 is able to account for global nonstationarities across land/ocean boundaries, as well as across mountain ranges. Inference can be achieved by balancing memory storage and distributed computation for a big data set of 220 million points. Once fitted with as few as five runs, the statistical model can generate surrogates fast and efficiently on a simple laptop, and provide uncertainty assessments very close to those obtained from all the available climate simulations (forty) on a monthly scale.

Key words: Big data; Nonstationarity; Spatio-temporal covariance model; Sphere; Stochastic generator; Tukey g-and-h autoregressive model; Wind Energy.

Short title: Stochastic Monthly Wind Generators

[1]Department of Mathematics and Statistics, 5752 Neville Hall, University of Maine, Orono, ME 04469, USA. E-mail: jaehong.jeong@maine.edu.

[2]Statistics Program, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia. E-mail: yuan.yan@kaust.edu.sa, marc.genton@kaust.edu.sa.

[3]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.

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.

1 Introduction

Wind energy plays an important role in many countries' energy portfolio as a significant renewable source 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 trajectories of greenhouse gases concentration (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 for assessing 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 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, an abbreviation for 'Stochastic Generator of Climate Model Output'[1], acts as a stochastic approximation of the climate model and hence

[1]not to be confounded with a Stochastic Weather Generator, which is focused on in-situ data at high temporal resolution

allows for sampling more surrogate climate runs[2]. In their study, the authors present a SG for the global annual wind and show that only five runs are sufficient to generate synthetic runs visually indistinguishable from the original simulations, and with a similar spatio-temporal local dependence. However, while the SG introduced by the authors is able to approximate annual global data for the Arabian peninsula effectively, an annual scale is not useful for wind energy assessment, and a SG at a finer temporal resolution in the same region is required to provide policy-relevant results.

A SG for monthly global wind output requires considerable modeling and computational efforts. From a modeling perspective, data indexed on the sphere and in time require a dependence structure able to incorporate complex nonstationarities across the entire Earth's system, see Jeong et al. (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 are particularly useful as they can provide flexible nonstationary structures for Gaussian processes in the spectral domain while maintaining 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 graphical descriptors such as land/ocean indicators and mountain ranges to be incorporated in a spatially varying spectrum.

Besides the modeling complexity, the computational challenges are remarkable, as inference needs to be performed on a big data set. Over the last two decades, the increase in the size of spatio-temporal data sets in climate has prompted the development of many new classes of scalable models. Among the many solutions proposed, fixed rank methods (Cressie and Johannesson, 2008), predictive processes (Banerjee et al., 2008), covariance tapering (Furrer et al., 2006), and Gaussian Markov random fields (Rue and Held, 2005) have played a key role in our ability to couple feasibility of the inference while retaining essential information to be communicated to

[2]A brief discussion on the difference between a SG and an emulator is contained in the same work.

stakeholders; see Sun et al. (2012) for a review. However, even by modern spatio-temporal data set standards, 220 million points is a considerable size, and to perform inference, a methodology that leverages on both the parallel computing and gridded geometry of the data is absolutely necessary. Castruccio and Genton (2018) have provided a framework for a fast and parallel methodology for big climate data sets. However, it has so far been limited to Gaussian processes. Whether an extension to non-Gaussian models with such a big data set is possible (and how) was an open question.

In this paper, we propose a SG for monthly winds that is multi-step, spectral and can capture a 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, 2018), defined as $Y(t) = \xi + \omega\tau_{g,h}\{Z(t)\}$, where ξ is a location parameter, ω 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):

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

where g controls the skewness and h governs the tail behavior. A significant advantage of Tukey g-and-h autoregressive processes is that they provide very flexible marginal distributions, allowing skewness and heavy tails to be adjusted. This class of non-Gaussian processes is integrated within the multi-step spectral scheme to still allow inference 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. The article ends with concluding remarks in Section 6.

2 The Community Earth System Model (CESM) Large ENSemble project (LENS)

We work on global wind data from LENS, which is an ensemble of CESM runs with version 5.2 of the Community Atmosphere Model from NCAR (Kay et al., 2015). The ensemble comprises runs at $0.9375° \times 1.25°$ (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 monthly near-surface wind speed at 10 m above the ground level (U10 variable) from 2006 to 2100. Since our focus is on future wind trends, we analyze the projections for a total of 95 years. We consider all 288 longitudes, and we discard latitudes near the poles to avoid numerical instabilities, due to the very close physical distance of neighboring points and the very different statistical behavior of wind speed in the Arctic and Antarctic regions (McInnes et al., 2011), and consistently with previous works. Therefore, we use 134 bands between 62°S and 62°N, and the full data set comprises approximately 220 million points ($5 \times 1140 \times 134 \times 288$). 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 Northern Tropic and latitude 60°N, the mean of wind speed over the ocean in September is stronger than that in March.

For each site, we test the significance of skewness and kurtosis of 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), thus indicating how the first two moments are not sufficient to characterize the temporal behavior of monthly wind in time. Most land points have significant skewness and, consistently with Bauer (1996), we observe that monthly wind speeds

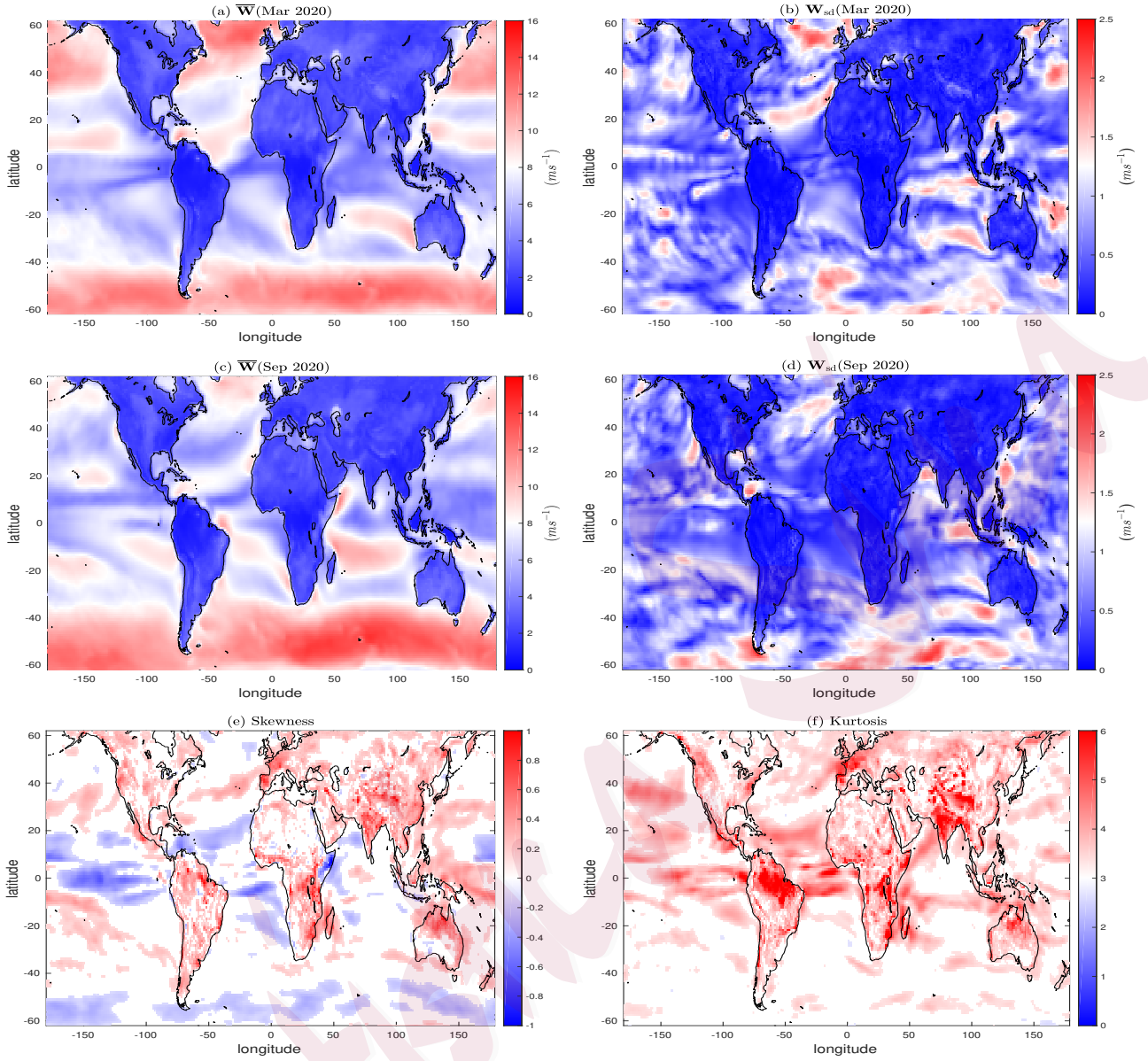

Figure 1: The (a) ensemble mean $\overline{\mathbf{W}}(\text{March 2020}) = \sum_{r=1}^{R} \mathbf{W}_r(\text{March 2020})/R$ where $R = 5$ is the number of ensemble members and (b) ensemble standard deviation $\mathbf{W}_{\text{sd}}(\text{March 2020}) = \sqrt{\sum_{r=1}^{R}\{\mathbf{W}_r(\text{March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

over the ocean are negatively skewed in the tropics and positively skewed outside of that region. The Tropical Indian Ocean and the Western Pacific Ocean, both areas of small wind speeds, represent an exception of 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, and to evolve randomly while still being attracted by the mean climate. Since 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 in this work. 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 ℓ_n, and time t_k, where $r = 1, \ldots, R$, $m = 1, \ldots, M$, $n = 1, \ldots, N$, and $k = 1, \ldots, K$, and define $\mathbf{W}_r = \{W_r(L_1, \ell_1, t_1), \ldots, W_r(L_M, \ell_N, t_K)\}^\top$.

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

$$\mathbf{D}_r = \underline{\boldsymbol{\xi}} + \underline{\boldsymbol{\omega}} \cdot \boldsymbol{\tau}_{\underline{\boldsymbol{g}},\underline{\boldsymbol{h}}}(\boldsymbol{\epsilon}_r), \quad \boldsymbol{\epsilon}_r \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}_{\text{space-time}})), \tag{2}$$

where $\underline{\boldsymbol{\xi}} = \boldsymbol{\xi} \otimes \mathbf{1}_K$, with $\boldsymbol{\xi} = \{\xi(L_1, \ell_1), \ldots, \xi(L_M, \ell_N)\}^\top$ being the vector of the location parameters, $\otimes$ the Kroneker product, $\mathbf{1}_K$ the vector of ones of length K, $\underline{\boldsymbol{\omega}} = \boldsymbol{\omega} \otimes \mathbf{1}_K$, with $\boldsymbol{\omega} = \{\omega(L_1, \ell_1), \ldots, \omega(L_M, \ell_N)\}^\top$ the vector of scale parameters, $\underline{\boldsymbol{g}} = \boldsymbol{g} \otimes \mathbf{1}_K$, and $\underline{\boldsymbol{h}} = \boldsymbol{h} \otimes \mathbf{1}_K$, with $\boldsymbol{g} = \{g(L_1, \ell_1), \ldots, g(L_M, \ell_N)\}^\top$ and $\boldsymbol{h} = \{h(L_1, \ell_1), \ldots, h(L_M, \ell_N)\}^\top$ the vectors of the MN parameters for the Tukey g-and-h transformation at each site. Here $\boldsymbol{\tau}_{\underline{\boldsymbol{g}},\underline{\boldsymbol{h}}}()$ represents the element-wise transformation according to (1), so that 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 denote by $\boldsymbol{\theta}_{\text{Tukey}} = \{\boldsymbol{\theta}_{\text{T};m,n}\}_{m,n}$, where $m = 1,\dots,M$, $n = 1,\dots,N$ and $\boldsymbol{\theta}_{\text{T};m,n} = \{\xi(L_m,\ell_n), \omega(L_m,\ell_n), g(L_m,\ell_n), h(L_m,\ell_n)\}^\top$ are the parameters of the Tukey g-and-h transformation and by $\boldsymbol{\theta}_{\text{space-time}} = (\boldsymbol{\theta}_{\text{time}}^\top, \boldsymbol{\theta}_{\text{lon}}^\top, \boldsymbol{\theta}_{\text{lat}}^\top)^\top$ 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 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) provided some 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 provide a description for 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 (2) evolves according to a Vector AutoRegressive model of order p (VAR(p)) with different parameters for each spatial location:

$$\boldsymbol{\epsilon}_r(t_k) = \boldsymbol{\Phi}_1\boldsymbol{\epsilon}_r(t_{k-1}) + \cdots + \boldsymbol{\Phi}_p\boldsymbol{\epsilon}_r(t_{k-p}) + \mathbf{S}\mathbf{H}_r(t_k), \quad \mathbf{H}_r(t_k) \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{C}(\boldsymbol{\theta}_{\text{lon}}, \boldsymbol{\theta}_{\text{lat}})), \tag{3}$$

where $\boldsymbol{\Phi}_1 = \text{diag}\{\phi^1_{L_m,\ell_n}\}, \dots, \boldsymbol{\Phi}_p = \text{diag}\{\phi^p_{L_m,\ell_n}\}$ are $MN \times MN$ diagonal matrices with autoregressive coefficients, and $\mathbf{S} = \text{diag}\{S_{L_m,\ell_n}\}$ is the diagonal matrix of the standard deviations.

Such a model assumes that there is no cross-temporal dependence across locations, and 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, 2018). The vector of temporal parameters is therefore $\boldsymbol{\theta}_{\text{time}} = \{\boldsymbol{\theta}_{\text{t};m,n}\}_{m,n}$, where $\boldsymbol{\theta}_{\text{t};m,n} = (\phi^1_{L_m,\ell_n}, \ldots, \phi^p_{L_m,\ell_n}, S_{L_m,\ell_n})$. The vectors $\boldsymbol{\theta}_{\text{Tukey}}$ and $\boldsymbol{\theta}_{\text{time}}$ in (2) are estimated simultaneously via a maximum approximated likelihood estimation (MALE, Xu and Genton (2015)), and since the model assumes no cross-temporal dependence, $\boldsymbol{\theta}_{\text{T};m,n}$ and $\boldsymbol{\theta}_{\text{t};m,n}$ can be estimated independently and in parallel across m and n.

Exact likelihood inference for the Tukey g-and-h autoregressive process is computationally expensive because the inverse Tukey g-g transformation $\tau^{-1}_{g,h}$ does not have an explicit form (except when either g or h is equal to 0). The idea of the MALE is to approximate $\tau^{-1}_{g,h}$ by a piecewise linear function $\tilde{\tau}^{-1}_{g,h}$, which would reduce the computational time considerably compared to calculating $\tau^{-1}_{g,h}$ numerically for each iteration in the optimization. The approximated log-likelihood function $\tilde{l}$ of the monthly residual wind speed time series, $\mathbf{D}_r(L_m, \ell_n) = \{D_r(L_m, \ell_n, t_1), \ldots, D_r(L_m, \ell_n, t_K)\}$, from ensemble r at latitude L_m and longitude ℓ_n can be written as:

$$\begin{aligned}\tilde{l}(\boldsymbol{\theta}_{\text{T};m,n}, \boldsymbol{\theta}_{\text{t};m,n} \mid \mathbf{D}_r(L_m, \ell_n),) =& f_{m,n}(\epsilon_1) + f_{m,n}(\epsilon_2|\epsilon_1) + \cdots + f_{m,n}(\epsilon_K|\epsilon_{K-1}, \ldots, \epsilon_{K-p}) \\ &- K\log\{\omega(L_m, \ell_n)\} - \frac{h(L_m, \ell_n)}{2}\sum_{k=1}^{K}\epsilon_k^2 \\ &- \sum_{k=1}^{K}\log\left(\exp\{g(L_m, \ell_n)\epsilon_k\} + \frac{h(L_m, \ell_n)}{g(L_m, \ell_n)}[\exp\{g(L_m, \ell_n)\epsilon_k\} - 1]\epsilon_k\right),\end{aligned} \tag{4}$$

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

For each location and each ensemble, the MALE is obtained by maximizing (4) with the optimal order p selected by BIC. The results are shown in Figure S3(b). For a substantial share

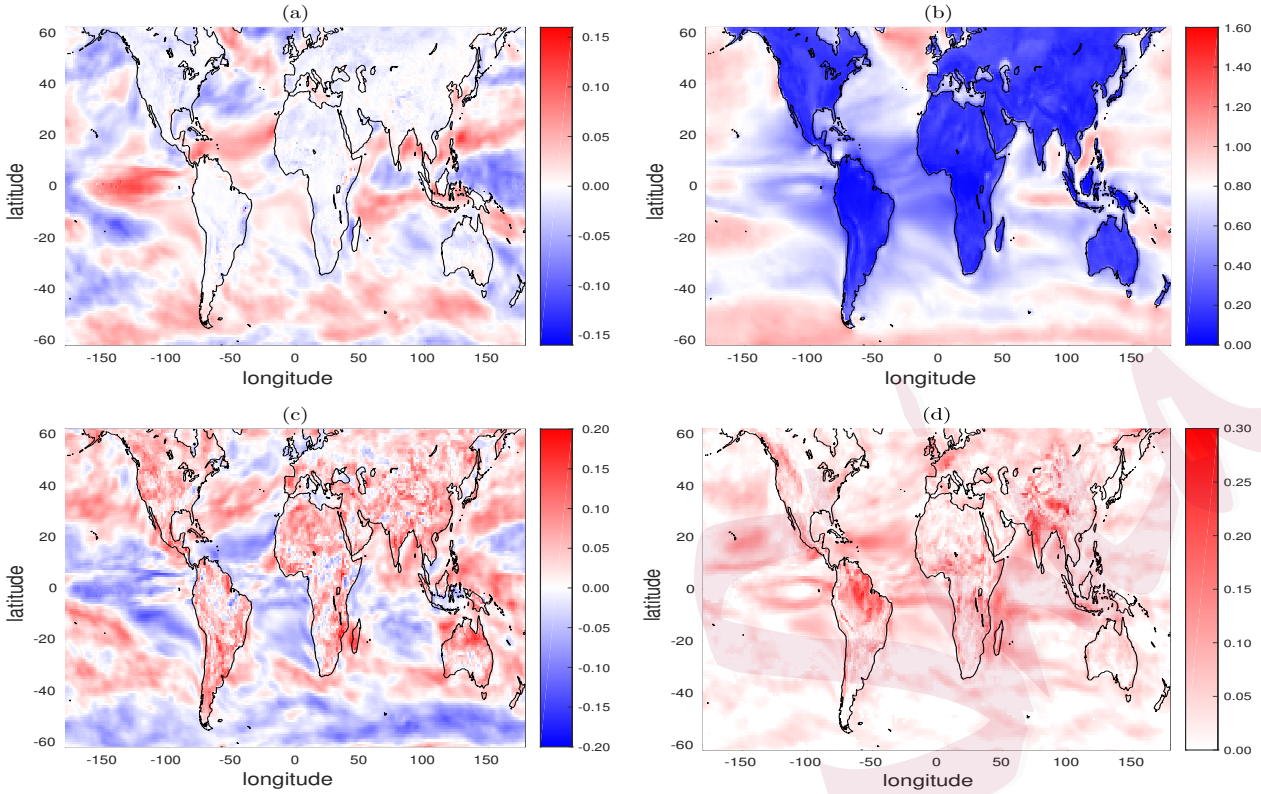

Figure 2: Plot of the estimated parameters $\hat{\boldsymbol{\theta}}_{\text{Tukey}}$ for the Tukey g-and-h transformation, (a) location, (b) scale, (c) g, and (d) h.

of points (56.1%), $p > 0$ was selected, hence underscoring the need for a model with temporal dependence, even after differencing the original data from the average across realizations. A map of $\hat{\phi}^1_{L_m,\ell_n}, \hat{\phi}^2_{L_m,\ell_n}$ and $\hat{\phi}^3_{L_m,\ell_n}$ is shown in the supplementary material (Figure S4), along with the p–values (Figure S5).

Estimated values $\hat{\boldsymbol{\theta}}_{\text{Tukey}}$ are represented in Figure 2. Here $\hat{g}(L_m, \ell_n)$ and $\hat{h}(L_m, \ell_n)$ were estimated with significant non-zero 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,

$$\hat{H}_r(L_m, \ell_n, t_k) = \frac{1}{\hat{S}(L_m, \ell_n)} \left\{ \hat{\epsilon}_r(L_m, \ell_n, t_k) - \hat{\phi}^1_{L_m,\ell_n} \hat{\epsilon}_r(L_m, \ell_n, t_{k-1}) - \cdots - \hat{\phi}^p_{L_m,\ell_n} \hat{\epsilon}_r(L_m, \ell_n, t_{k-p}) \right\}, \quad (5)$$

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

are the inverse Tukey g-and-h transformations at latitude L_m and longitude ℓ_n.

The following sections provide a model for the dependence structure of $\mathbf{H}_r(t_k)$, i.e., a parametrization of $\mathbf{C}(\boldsymbol{\theta}_{\text{lon}}, \boldsymbol{\theta}_{\text{lat}})$ in (3). Specifying a valid model for the entire spherical domain that is able to capture global dependence structures is a non-trivial task. However, the following steps rely on the Gaussianity of $\mathbf{H}_r(t_k)$, and hence require to specify only the covariance structure.

3.3 Step 2: Longitudinal Structure

Here, we focus on $\boldsymbol{\theta}_{\text{lon}}$, i.e., we provide a model for the dependence structure at different longitudes but at the same latitude. Since 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 successfully implemented 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 we model $H_r(L_m, \ell_n, t_k)$ in the spectral domain via a generalized Fourier transform across longitude. Indeed, if we define $\iota = \sqrt{-1}$ to be the imaginary unit, $c = 0, \dots, N-1$ the wavenumber, then the process can be spectrally represented as

$$H_r(L_m, \ell_n, t_k) = \sum_{c=0}^{N-1} f_{L_m,\ell_n}(c) \exp(\iota \ell_n c) \widetilde{H}_r(c, L_m, t_k), \tag{6}$$

with $f_{L_m,\ell_n}(c)$ being a spectrum evolving across longitude, and $\widetilde{H}_r(c, L_m, t_k)$ 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 (consistently with Jeong et al., 2018) are treated as covariates. Therefore $f_{L_m,\ell_n}(c)$ depends on ℓ_n being in a land, ocean, and

high mountain domain, with the following expression:

$$f_{L_m,\ell_n}(c) = \begin{cases} f^1_{L_m,\ell_n}(c) & \text{if } (L_m,\ell_n) \in \text{high mountain}, \\ f^2_{L_m,\ell_n}(c) b_{\text{land}}(L_m,\ell_n; g'_{L_m}, r'_{L_m}) & \text{if } (L_m,\ell_n) \in \text{land}, \\ f^3_{L_m,\ell_n}(c)\{1 - b_{\text{land}}(L_m,\ell_n; g'_{L_m}, r'_{L_m})\} & \text{if } (L_m,\ell_n) \in \text{ocean}, \end{cases} \tag{7}$$

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 the land and the ocean domain. Each of the three components of the spectrum in (7) is parametrized by (Castruccio and Stein, 2013; Poppick and Stein, 2014): $|f^j_{L_m,\ell_n}(c)|^2 = \psi^j_{L_m,\ell_n}\{(\alpha^j_{L_m,\ell_n})^2 + 4\sin^2(c\pi/N)\}^{-\nu^j_{L_m,\ell_n}-1/2}$, for $j = 1,2,3$, where $(\psi^j_{L_m,\ell_n}, \alpha^j_{L_m,\ell_n}, \nu^j_{L_m,\ell_n})$ are interpreted as the variance, inverse range, and smoothness parameters, respectively, similarly as for the Matérn spectrum. The parameters are modeled so that their logarithm changes continuously and linearly depends on the altitude, i.e., $\psi^j_{L_m,\ell_n} = \beta^{j,\psi}_{L_m} \exp[\tan^{-1}\{A_{L_m,\ell_n}\gamma^{\psi}_{L_m}\}]$, $j = 1,2$ and $\psi^3_{L_m,\ell_n} = \beta^{3,\psi}_{L_m}$, where $\beta^{j,\psi}_{L_m} > 0$, $\gamma^{\psi}_{L_m} \in \mathbb{R}$, and A_{L_m,ℓ_n} is the altitude at location (L_m,ℓ_n). Similar notation holds for $\alpha^j_{L_m,\ell_n}$ and $\nu^j_{L_m,\ell_n}$. Hence, the longitudinal parameters are $\boldsymbol{\theta}_{\text{lon}} = \{\boldsymbol{\theta}_{\ell,m}\}_m$, where $\boldsymbol{\theta}_{\ell,m} = \{(\beta^{j,\psi}_{L_m}, \gamma^{\psi}_{L_m}, \beta^{j,\alpha}_{L_m}, \gamma^{\alpha}_{L_m}, \beta^{j,\nu}_{L_m}, \gamma^{\nu}_{L_m}, g'_{L_m}, r'_{L_m})^\top, j = 1,2,3\}$. Since the $\boldsymbol{\theta}_{\ell,m}$ are independent across m, model inference across latitude can be performed independently with distributed computing.

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

$$\begin{aligned} 2l(\boldsymbol{\theta}_{\ell,m} \mid \mathbf{H}) &= KN(R-1)\log(2\pi) + KN\log(R) \\ &\quad + (R-1)K\log|\boldsymbol{\Sigma}(\boldsymbol{\theta}_{\ell,m})| + \sum_{k=1}^{K}\sum_{r=1}^{R} \mathbf{H}_r^\top(t_k)\boldsymbol{\Sigma}(\boldsymbol{\theta}_{\ell,m})\mathbf{H}_r(t_k), \end{aligned} \tag{8}$$

where $\boldsymbol{\Sigma}(\boldsymbol{\theta}_{\ell,m})$ is the $N \times N$ covariance matrix of latitudinal band L_m as implied by (6) and (7).

3.4 Step 3: Latitudinal Structure

We now provide a model for latitudinal dependence, and since the model in (6) is independent and identically distributed across r and t_k, we take out these two indices for simplicity. While Castruccio and Stein (2013) and later works have proposed an autoregressive model for $\widetilde{H}(c, L_m)$ across m (but independent across c), we consider a more general Vector AutoRegressive model of order 1 (VAR(1)) so that $\widetilde{H}(c, L_m)$ is allowed to also depend on neighboring wavenumbers. We define $\widetilde{\mathbf{H}}_{L_m} = \{\widetilde{H}(1, L_m), \dots, \widetilde{H}(N, L_m)\}^\top$ and the latitudinal dependence by $\widetilde{\mathbf{H}}_{L_m} = \boldsymbol{\varphi}_{L_m}\widetilde{\mathbf{H}}_{L_{m-1}} + \mathbf{e}_{L_m}$, where $\mathbf{e}_{L_m} \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{L_m})$, and $\boldsymbol{\varphi}_{L_m}$ is a matrix of size $N \times N$ with 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 ten sequential sub-samples of 95 years (10 years except for the last partition) to reduce the computation burden. We derive an expression similar to (8) for this step, and estimate ζ_{L_m} and η_{L_m} from each of the 10 sub-samples, as shown in Figure S7 (other estimates of longitudinal dependence parameters show similar patterns). Since there is no evidence of a change in latitudinal dependence over time, we consider the average of parameter estimates. Such value is used for combining multiple latitudinal bands and generating 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

Inference for a dataset indexed in latitude, longitude, time, and realization comprising of 220 million data points is a daunting task even with the aforementioned stepwise approach, which

allows 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 on the gridded geometry of the data. Indeed, (8) can be expressed equivalently in the spectral domain through a fast Fourier transform of the data (Whittle likelihood, Whittle (1953)), so that the computational complexity is reduced from $\mathrm{O}(M^3N^3)$ to $\mathrm{O}(M^2N\log N)$ and storage from $\mathrm{O}(M^2N^2)$ to $\mathrm{O}(M^2N)$.

We used a workstation with 2×12 cores Intel Xeon E5-2680V3 2.5GHz. Step 1 required approximately 6 hours, step 2 required 29 hours, and step 3 required 179 hours, for a total of approximately nine days. While inference is therefore nontrivial and requires considerable computational resources, once the parameters are estimated, generation of forty statistical surrogates as required for Section 5 required only 16 minutes on a simple laptop (see the Matlab Graphical Users 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 two models with special cases of spatial dependence structure from steps 2 and 3 detailed in Sections 3.3 and 3.4. In the supplementary material, we provide additional comparison 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, we notice that the G-AR process can be obtained from (2) 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 Bayesian Information Criterion (BIC) between the two models at each site from one ensemble member. Positive and negative values indicate better and worse model fit of TGH-AR compared to G-AR, respectively. TGH-AR outperforms G-AR in more than 85% of spatial

locations, with a considerable improvement in the BIC score (the map scale is in the order of 10^3). The fit for land sites is overall considerably better for TGH-AR, with 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.

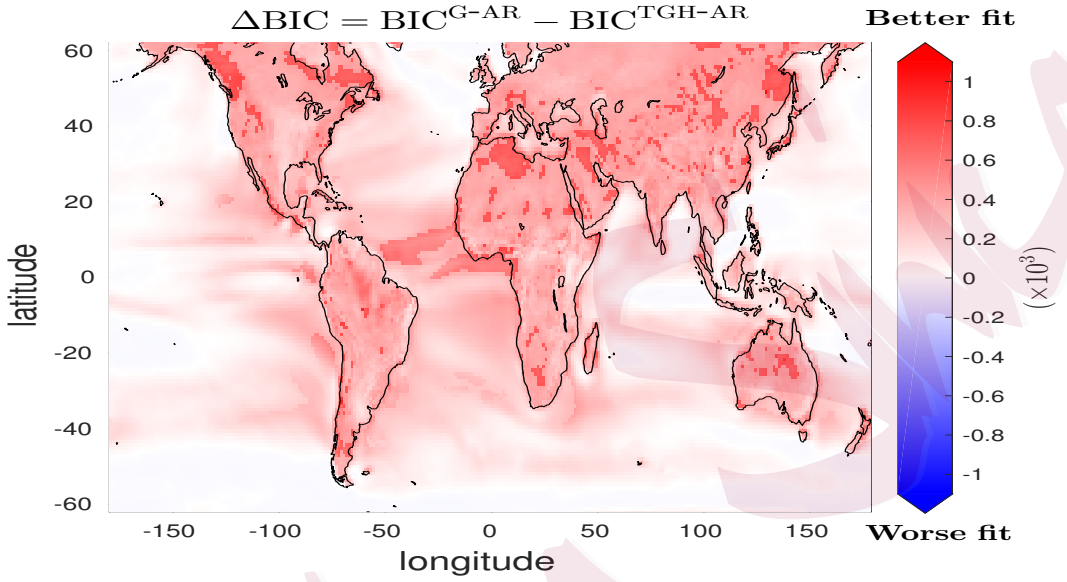

Figure 3: Map of differences in BIC between TGH-AR and G-AR from one ensemble member.

4.2 Comparison with sub-models of global dependence

The TGH-AR model is also compared with one of no altitude dependence, i.e., where

$$\psi^1_{L_m,\ell_n} = \psi^2_{L_m,\ell_n}, \alpha^1_{L_m,\ell_n} = \alpha^2_{L_m,\ell_n}, \nu^1_{L_m,\ell_n} = \nu^2_{L_m,\ell_n} \implies f^1_{L_m,\ell_n}(c) = f^2_{L_m,\ell_n}(c)$$

for all m, n, c in (7). 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 TGH-AR with a model having an autoregressive dependence across latitude, i.e., a model in which $a_{L_m} = b_{L_m} = 0$ in the parametrization of $\boldsymbol{\varphi}_{L_m}$ in Section 3.4, which we denote as ARL.

Since both LAO and ARL are 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 LAO to ARL, hence the need to incorporate the altitude while modeling the

covariance structure. The additional, smaller (although non-negligible, as the BIC improvement is approximately 10^5) improvement from ARL to TGH-AR underscores the necessity of a flexible model that is able to account for dependence across both wavenumbers and latitude.

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 $\Delta\text{loglik}/\{NMK(R-1)\}$ are that values above 0.1 are considered to be large and those above 0.01 are modest but still sizable (Castruccio and Stein, 2013).

Model	LAO	ARL	TGH-AR
# of parameters	1338	2142	2408
$\Delta\text{loglik}/\{NMK(R-1)\}$	0	0.0440	**0.0443**
BIC ($\times 10^8$)	-5.8963	-6.0511	$\mathbf{-6.0521}$

All three models can also be compared via local contrasts, as the residuals in (5) 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):

$$\begin{aligned}\Delta_{ew;m,n} &= \frac{1}{KR}\sum_{k=1}^{K}\sum_{r=1}^{R}\{H_r(L_m,\ell_n,t_k)-H_r(L_m,\ell_{n-1},t_k)\}^2,\\ \Delta_{ns;m,n} &= \frac{1}{KR}\sum_{k=1}^{K}\sum_{r=1}^{R}\{H_r(L_m,\ell_n,t_k)-H_r(L_{m-1},\ell_n,t_k)\}^2,\end{aligned} \tag{9}$$

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

We compare ARL with 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

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

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

in the case of the east-west variances. A representation of these differences for the small region of interest near South Africa (13.75°E $\sim$ 48.75°E and 30°S $\sim$ 4°N) is given in Figure S11. Positive values are obtained when TGH-AR is a better model fit than the ARL; negative values are obtained when ARL is the better model fit. Figure S11(a) and (b) show that dark red colors are more widely spread over mountains, and that no clear difference is shown over the ocean. Results presented in Table 2 are consistent with the visual inspection and the two metrics, in particular, over mountain 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 obtain an assessment of the uncertainty of monthly wind power density, 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 estimate is highly variable. For each latitude and longitude (i.e., each n and m), we fit a spline $\widetilde{W}(L_m, \ell_n, t_k)$ which minimizes the following function (Castruccio and Guinness, 2017; Jeong et al., 2018): $\lambda \sum_{k=1}^{K} \left\{ \overline{W}(L_m, \ell_n, t_k) - \widetilde{W}(L_m, \ell_n, t_k) \right\}^2 + (1-\lambda) \sum_{k=1}^{K} \left\{ \nabla_2 \widetilde{W}(L_m, \ell_n, t_k) \right\}^2$, ∇_2 being 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, 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 non-parametric 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 easily generated by the Algorithm 1.

Algorithm 1 Generate surrogates

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

We generate forty SG runs with the model presented in this work and compare them with the original forty 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 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 et al., 2018). We observe that the SG runs from the Tukey g-and-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 LENS and SG in terms of 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 near-future wind speed trends near the Indian ocean for each of the SG and LENS runs. 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.

We subsequently provide an assessment of the wind energy potential. The wind power density (WPD) (in Wm^{-2}) evaluates the wind energy resource available at the site for conversion by a wind turbine. WPD can be calculated as $\text{WPD} = 0.5\rho u^3, u = u_r(z/z_r)^\alpha$, where ρ is the air density ($\rho = 1.225$ kgm^{-3} in this study), u is the wind speed at a certain height z, u_r is the

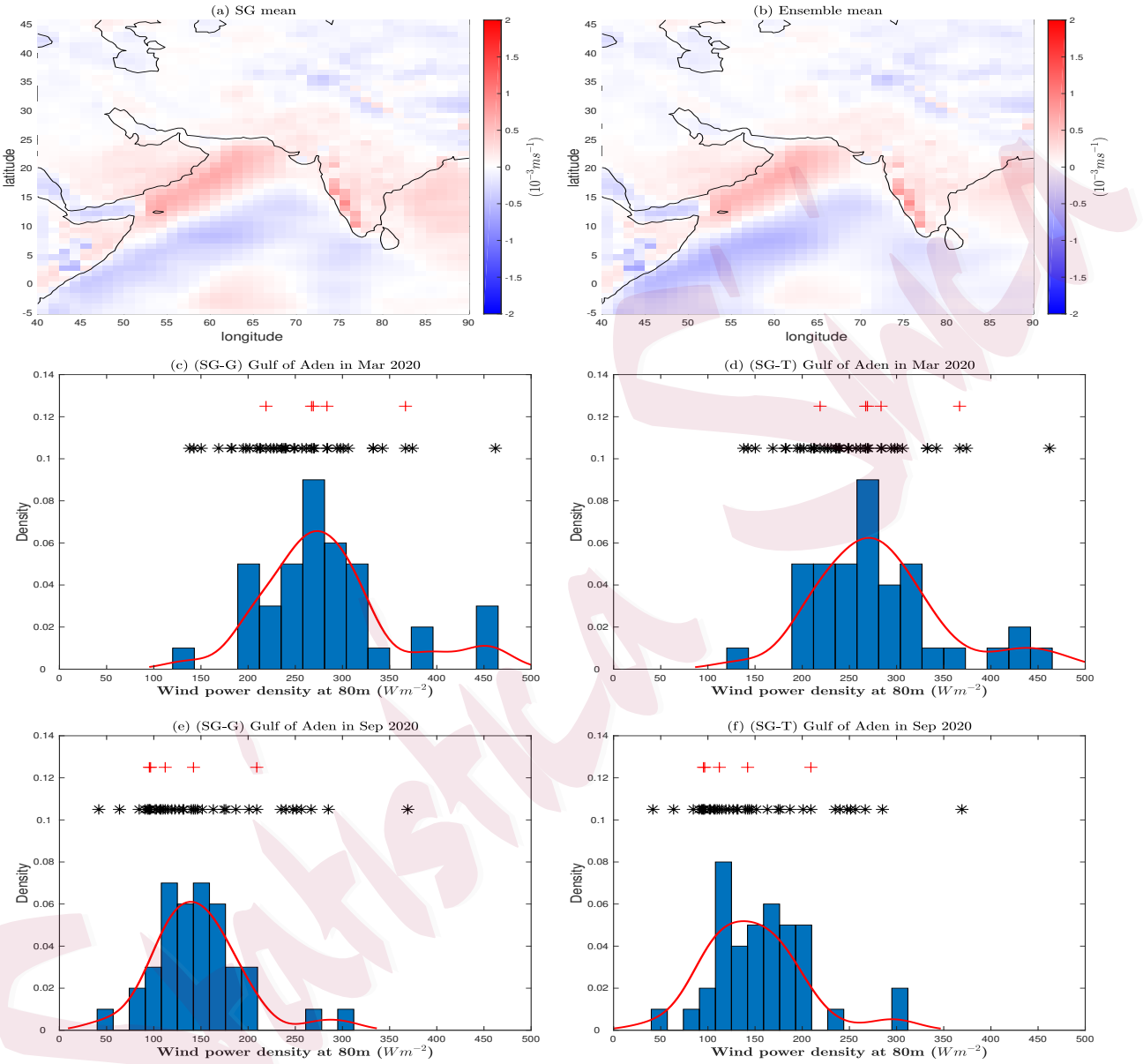

Figure 4: Maps of (a) the mean from SG runs and (b) the ensemble mean from the near-future (2013–2046) near-surface wind speed trends near the Indian ocean. Histogram of the distribution of the wind power density at 80 m with nonparametric density in red for the forty SG-G and SG-T runs near the Gulf 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).

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°E and 12.72°N), a narrow channel connecting the Red Sea to the Indian Ocean with high wind regimes (Yip et al., 2017), and we choose to work on WPD at 80 meters, a standard height for wind turbines (Holt and Wang, 2012; Yip et al., 2017), in the year 2020.

For completeness, we also considered the Gaussian-based SG runs. Now we call the Gaussian-based SG runs as the SG-G runs, and our original SG runs throughout the work are called the SG-T runs to distinguish these two SG runs. Results for March and September 2020 are represented in Figure 4(c,d) and (e,f), with the histograms representing both SG-G and 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 rightly skewed distributional shapes, as also reflected by 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), and matches the uncertainty generated by the forty LENS runs. Figure S17 shows the comparison on 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 report the boxplots of the distribution of 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 a slight misfit in April and November. The importance of such results cannot be understated: both SG runs are able to capture the interannual WPD patterns, as well as 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), and the SGs can reproduce the same range with as little

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 from the 40 LENS runs than the SG-G case. We computed the differences of the skewness and kurtosis values between the SG runs and LENS runs at each month in 2020, and then 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. Also, the metric in the kurtosis values for both cases were 0.8926 (0.5586) and 0.7948 (0.4531), respectively.

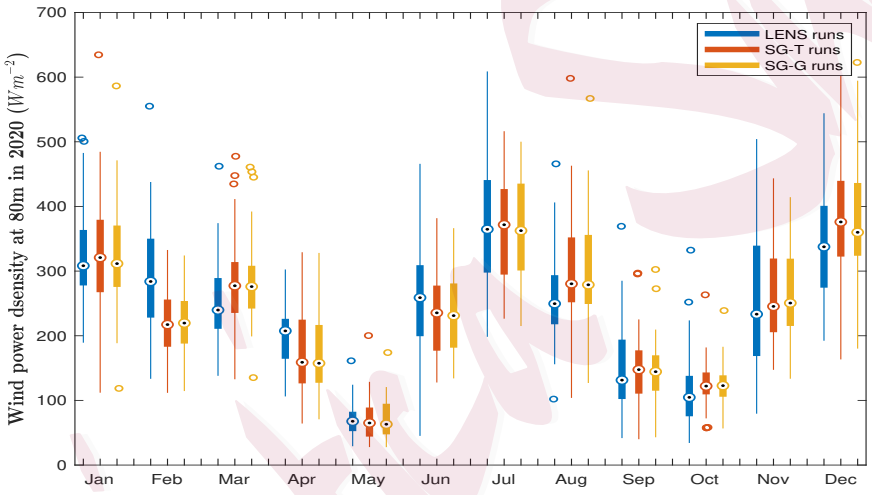

Figure 5: Boxplots of the distribution of the wind power density at 80 m, in 2020, for 40 LENS runs and the 40 SG runs based on the Tukey g-and-h and Gaussian cases near the Gulf of Aden.

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 Users 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 it requires only the stored estimated parameters, along with the algorithm described in this section for data generation, approximately 123 Mb to generate as many ensembles as desired, while 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 for approximating a computer output with a faster surrogate, we provided a fast, parallelizable and scalable methodology to perform inference 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 controlling skewness and tail behavior with two distinct parameters. This class of models is embedded in a multi-step approach to allow inference for a nonstationary global model while also capturing site-specific temporal dependence, and it clearly outperforms currently available Gaussian models.

Our model has been applied as a SG, a new class of stochastic approximations that assesses more efficiently the internal variability for wind energy resources in developing countries with poor observational data coverage, using global models. Our results suggest that the uncertainty produced by the SG with a training set of five runs is very similar to that from the forty 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 towards the use of SGs at policy-relevant time scales.

While we focused on global wind energy assessment, 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 other datasets not related to geoscience, whenever the data suggests different scales of spatio-temporal dependence. As an example, Castruccio et al. (2018) applied a similar step-wise approach to fMRI data, with spatial dependence both at voxel, regional and whole-brain level.

References

Archer, C. L. and M. Z. Jacobson (2003). Spatial and Temporal Distributions of US Winds and Wind Power at 80 m Derived From Measurements. *Journal of Geophysical Research: Atmospheres 108*(D9). 4289.

Bai, J. and S. Ng (2005). Tests for Skewness, Kurtosis, and Normality for Time Series Data. *Journal of Business & Economic Statistics 23*(1), 49–60.

Banerjee, S., A. E. Gelfand, A. O. Finley, and H. Sang (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*(10), 1087–1102.

Castruccio, S. and M. G. Genton (2016). Compressing an Ensemble with Statistical Models: An Algorithm for Global 3D Spatio-Temporal Temperature. *Technometrics 58*(3), 319–328.

Castruccio, S. and M. G. Genton (2018). Principles for Inference on Big Spatio-Temporal Data from Climate Models. *Statistics and Probability Letters 136*, 92–96.

Castruccio, S., M. G. Genton, and Y. Sun (2018). Visualising Spatio-Temporal Models with Virtual Reality: From Fully Immersive Environments to Apps in Stereoscopic View. *Journal of the Royal Statistical Society - Series A (with discussion)*. in press.

Castruccio, S. and J. Guinness (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*(2), 329–344.

Castruccio, S., H. Ombao, and M. G. Genton (2018). A Scalable Multi-Resolution Spatio-Temporal Model for Brain Activation and Connectivity in fMRI Data. *Biometrics 74*(3), 823–833.

Castruccio, S. and M. L. Stein (2013). Global Space-Time Models for Climate Ensembles. *The Annals of Applied Statistics 7*(3), 1593–1611.

Cressie, N. and G. Johannesson (2008). Fixed Rank Kriging for Very Large Spatial Data Sets. *Journal of the Royal Statistical Society: Series B 70*(1), 209–226.

Davis, P. J. (1979). *Circulant Matrices*. American Mathematical Society.

Furrer, R., M. G. Genton, and D. Nychka (2006). Covariance Tapering for Interpolation of Large Spatial Datasets. *Journal of Computational and Graphical Statistics 15*(3), 502–523.

Holt, E. and J. Wang (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*(12), 2188–2202.

Jeong, J., S. Castruccio, P. Crippa, and M. G. Genton (2018). Reducing Storage of Global Wind Ensembles with Stochastic Generators. *The Annals of Applied Statistics 12*(1), 490–509.

Jeong, J., M. Jun, and M. G. Genton (2017). Spherical Process Models for Global Spatial Statistics. *Statistical Science 32*(4), 501–513.

Jun, M. and M. L. Stein (2008). Nonstationary Covariance Models for Global Data. *The Annals of Applied Statistics 2*(4), 1271–1289.

Kay, J. E., C. Deser, A. Phillips, A. Mai, C. Hannay, G. Strand, J. M. Arblaster, S. C. Bates, G. Danabasoglu, J. Edwards, M. Holland, P. Kushner, J.-F. Lamarque, D. Lawrence, K. Lindsay, A. Middleton, E. Munoz, R. Neale, K. Oleson, L. Polvani, and M. Vertenstein (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*(8), 1333–1349.

Lorenz, E. N. (1963). Deterministic Nonperiodic Flow. *Journal of the Atmospheric Sciences 20*(2), 130–141.

McInnes, K. L., T. A. Erwin, and J. M. Bathols (2011). Global Climate Model Projected Changes in 10 m Wind Speed and Direction due to Anthropogenic Climate Change. *Atmospheric Science Letters 12*(4), 325–333.

Newman, J. and P. Klein (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, Austin, TX*, Volume 7.

Obama, B. (2017). The Irreversible Momentum of Clean Energy. *Science*.

Peterson, E. W. and J. P. Hennessey Jr (1978). On the Use of Power Laws for Estimates of Wind Power Potential. *Journal of Applied Meteorology 17*(3), 390–394.

Poppick, A. and M. L. Stein (2014). Using Covariates to Model Dependence in Nonstationary, High-Frequency Meteorological Processes. *Environmetrics 25*(5), 293–305.

Rue, H. and L. Held (2005). *Gaussian Markov Random Fields: Theory and Applications*. Boca Raton, FL: CRC Press.

Sun, Y., B. Li, and M. G. Genton (2012). Geostatistics for Large Datasets. In E. Porcu, J. M. Montero, and M. Schlather (Eds.), *Advances and Challenges in Space-Time Modelling of Natural Events*, pp. 55–77. Springer.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

van Vuuren, D. P., J. Edmonds, M. Kainuma, K. Riahi, A. Thomson, K. Hibbard, G. C. Hurtt, T. Kram, V. Krey, J.-F. Lamarque, T. Masui, M. Meinshausen, N. Nakicenovic, S. J. Smith, and S. K. Rose (2011). The Representative Concentration Pathways: An Overview. *Climatic Change 109*, 5–31.

Wang, Z., A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli (2004). Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing 13*(4), 600–612.

Whittle, P. (1953). Estimation and Information in Stationary Time Series. *Arkiv för Matematik 2*, 423–434.

Wiser, R., Z. Yang, M. Hand, O. Hohmeyer, D. Infield, P. H. Jensen, V. Nikolaev, M. O'Malley, G. Sinden, and A. Zervos (2011). Wind Energy. In O. Edenhofer, R. Pichs-Madruga, Y. Sokona, K. Seyboth, P. Matschoss, S. Kadner, T. Zwickel, P. Eickemeier, G. Hansen, and S. Schlömer (Eds.), *IPCC Special Report on Renewable Energy Sources and Climate Change Mitigation*, pp. 535–608. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Xu, G. and M. G. Genton (2015). Efficient Maximum Approximated Likelihood Inference for Tukey's g-and-h Distribution. *Computational Statistics & Data Analysis 91*, 78–91.

Yan, Y. and M. G. Genton (2018). Non-Gaussian Autoregressive Processes with Tukey g-and-h Transformations. *Environmetrics 29*, e2503.

Yip, C. M. A., U. B. Gunturu, and G. L. Stenchikov (2017). High-Altitude Wind Resources in the Middle East. *Scientific Reports 7*(1), 9885.