The Annals of Applied Statistics
2018, Vol. 12, No. 1, 490–509
https://doi.org/10.1214/17-AOAS1105
© Institute of Mathematical Statistics, 2018

REDUCING STORAGE OF GLOBAL WIND ENSEMBLES WITH STOCHASTIC GENERATORS[1]

BY JAEHONG JEONG*, STEFANO CASTRUCCIO†, PAOLA CRIPPA† AND MARC G. GENTON*

King Abdullah University of Science and Technology and University of Notre Dame†*

Wind has the potential to make a significant contribution to future energy resources. Locating the sources of this renewable energy on a global scale is however extremely challenging, given the difficulty to store very large data sets generated by modern computer models. We propose a statistical model that aims at reproducing the data-generating mechanism of an ensemble of runs via a Stochastic Generator (SG) of global annual wind data. We introduce an evolutionary spectrum approach with spatially varying parameters based on large-scale geographical descriptors such as altitude to better account for different regimes across the Earth's orography. We consider a multi-step conditional likelihood approach to estimate the parameters that explicitly accounts for nonstationary features while also balancing memory storage and distributed computation. We apply the proposed model to more than 18 million points of yearly global wind speed. The proposed SG requires orders of magnitude less storage for generating surrogate ensemble members from wind than does creating additional wind fields from the climate model, even if an effective lossy data compression algorithm is applied to the simulation output.

1. Introduction. Environmental and societal concerns about climate change are prompting many countries to seek alternative energy resources [Moomaw et al. (2011), Obama (2017)]. Wind is a clean and renewable energy source that has the potential to substantially contribute to energy portfolios without causing negative environmental impacts [Wiser et al. (2011)] and that can reduce the quantity of anthropogenic greenhouse gases on global warming [Barthelmie and Pryor (2014)]. In order to provide energy assessments in developing countries where no regional studies are available, Earth System Models (ESMs) currently represent a valuable tool to investigate where sustainable wind resources are located. While ESMs are important for physically consistent projections, they represent only an approximation of the true state of the Earth's system, thereby representing uncertainty. In

Received June 2017; revised September 2017.

[1] 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.

Key words and phrases. Axial symmetry, nonstationarity, spatio-temporal covariance model, sphere, stochastic generator, surface wind speed.

particular, small perturbations in the initial conditions generate a plume of simulations whose uncertainty (internal variability) needs to be quantified. While performing sensitivity analysis from internal variability is a fundamental task, a typical collection (ensemble) of runs, such as the Coupled Model Intercomparison Phase 5 (CMIP5) [Taylor, Stouffer and Meehl (2012)], comprises a small number of ESM runs, making a detailed assessment infeasible. The Community Earth System Model (CESM) Large ENSemble project (LENS) from the National Center for Atmospheric Research (NCAR) was implemented to provide a large collection of climate model simulations to assess projections in the presence of internal variability with the same forcing scenario [Kay et al. (2015)]. This ensemble required an enormous effort for only a single scenario (10 million CPU hours and more than 400 terabytes of storage), and very few academic institutions or national research centers have the resources for such an undertaking.

To mitigate storage issues arising when generating such large amounts of data, NCAR has proposed a series of investigations on the topic of reducing storage needs for climate model output. Baker et al. (2014) investigated the applicability of lossless and lossy compression algorithms to climate model output. Lossless and lossy compression algorithms respectively provide an exact reconstruction of the data or a reconstruction with some loss of information. Baker et al. (2016) reported that a lossy algorithm for LENS achieves data reduction that does not impact general scientific conclusions. Guinness and Hammerling (2017) introduced a compression approach based on a set of summary statistics and a statistical model for the mean and covariance structure in the climate model output.

Statistical models can provide appropriate stochastic approximations of the spatio-temporal characteristics of the model output, and hence they can be used as surrogates of the original runs [Mearns et al. (2001)]. Castruccio and Stein (2013), Castruccio and Genton (2014), Castruccio and Genton (2016) and Castruccio and Guinness (2017) introduced a Stochastic Generator (SG) for annual temperature data to investigate internal variability for different ensembles, assuming that the observed ensemble members were realizations of an underlying statistical model. This approach allowed them to generate runs that were visually indistinguishable from the original model output. In this work, we operate under this framework.

This work is part of an ongoing collaborative effort with NCAR to develop solutions to deal with memory-intensive models and of a series of investigations sponsored by KAUST to develop novel statistical methodologies to assess wind resources in Saudi Arabia and more broadly in developing countries by relying on ESMs. Various approaches have been proposed to model wind in space and time; see the reviews by Soman et al. (2010) and Zhu and Genton (2012). For LENS, we establish a SG that accounts for the spatio-temporal dependence of the data and uses its parameters to generate additional surrogate runs and efficiently assess the uncertainty in multi-decadal projections.

Wind fields are expected to exhibit varying spatio-temporal smoothness across longitudes, which is associated with land/ocean regimes and orography. Differences in altitude produce thermal effects as well as acceleration of wind flows over hills, and funneling effects in narrow valleys [Banuelos-Ruedas, Angeles-Camacho and Rios-Marcuello (2011)], and these features are expected to impact the spatial smoothness of this variable. We introduce an evolutionary spectrum approach [Priestley (1965)],[2] coupled with spatially varying parameters depending on the surface altitude to better account for different regimes across the Earth's orography. We further introduce a model that allows the latitudinal spectral dependence to vary across different wavenumbers, which markedly improves the fit and allows to model complex latitudinal nonstationarities.

We perform inference via a multi-step conditional likelihood approach, and we show how the resulting model reduces computational burden and storage costs. Once the parameters are estimated, the proposed model can generate surrogates of ESM runs with different initial conditions within seconds on a modest laptop. The SG requires a small data set of approximately 30 megabytes that describes the mean structure and the parameters of the space–time covariance, whereas downloading a single wind variable from 40 LENS runs requires 1.1 gigabytes.

The remainder of the paper is organized as follows. Section 2 describes the LENS data set. Section 3 details the space–time statistical model and the inferential approach. Section 4 provides a model comparison and validation of local behavior. Section 5 illustrates how to generate runs, validate the large scale behavior and assess the internal variability of global wind fields and wind power densities. The article ends with Section 6, which offers a discussion and concluding remarks.

2. The large ensemble. We focus on LENS, an ensemble of CESM runs with version 5.2 of the Community Atmosphere Model from NCAR [Kay et al. (2015)]. The ensemble comprises 40 runs of coupled simulations for the period between 1920 and 2100 at $0.9375° \times 1.25°$ (latitude × longitude) resolution. Each member is subject to the same radiative forcing scenario: historical up to 2005 and the Representative Concentration Pathway (RCP) 8.5 [van Vuuren et al. (2011)] thereafter. We focus on yearly wind speed at 10 m (computed from the monthly U10 variable) and, since our focus is on future wind trends, we analyze the projections from 2006 to 2100, for a total of 95 years. In the supplementary material [Figure S1, Jeong et al. (2018)], we use a lack of fit index to assess the number of runs R required in the training set for a satisfactory fit, and for this work we establish $R = 5$, randomly chosen from the original ensemble. We consider all 288 longitudes, and we discard latitudes near the poles as they would lead to 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, Erwin and Bathols (2011)]. We therefore focus on 134 bands between 62°S and 62°N, and the full dataset comprises more than 18 million points ($5 \times 95 \times 134 \times 288$). In Figure 1, we show the ensemble mean and standard deviation of the yearly wind speed from the five chosen runs, in 2020.

[2] The evolutionary spectrum generalizes the spectrum of a stationary process, by allowing it to vary across longitude while still retaining positive definite covariance functions.

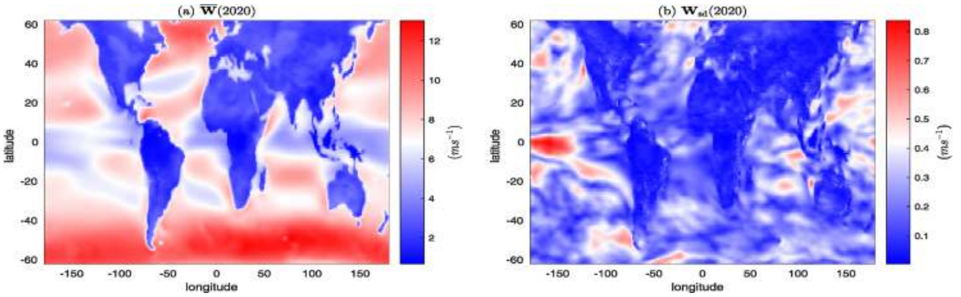

FIG. 1. *The* (a) *ensemble mean* $\overline{\mathbf{W}}(2020) = \sum_{r=1}^{R} \mathbf{W}_r(2020)/R$ *and* (b) *ensemble standard deviation* $\mathbf{W}_{\mathrm{sd}}(2020) = \sqrt{\sum_{r=1}^{R} \{\mathbf{W}_r(2020) - \overline{\mathbf{W}}(2020)\}^2/R}$, *where R is the number of ensemble members, of the yearly near-surface wind speed* (*in* ms^{-1}) *for* $R = 5$.

3. The space–time covariance model.

3.1. *A review of statistical models on a sphere.* Recently, Gneiting (2013) and Ma (2015) provided an overview of isotropic covariance functions for Gaussian processes on a sphere based on geodesic distance. Porcu, Bevilacqua and Genton (2016) proposed spatio-temporal covariance and cross-covariance models based on geodesic distance and Clarke, Alegría and Porcu (2016) studied the regularity properties of Gaussian random fields on a sphere across time. For nonstationary covariance models on a sphere, various construction approaches, such as differential operators [Jun and Stein (2007, 2008), Jun (2011, 2014)], spherical harmonic representation [Hitczenko and Stein (2012), Stein (2007)], stochastic partial differential equations [Bolin and Lindgren (2011), Lindgren, Rue and Lindström (2011)], kernel convolution [Heaton et al. (2014)] and deformation [Das (2000)] have been introduced. A new review of spherical process models for global spatial statistics can be found in Jeong, Jun and Genton (2017).

When modeling global data, a common assumption is that the (Gaussian) spatial process is *axially symmetric*, that is, its mean depends on latitude, L, and its covariance depends only on the longitudinal lag, $\ell_1 - \ell_2$, between two points [Jones (1963)]. This class of models implies that data are stationary at a given latitude, but this assumption is clearly inappropriate for many variables whose dynamics are influenced by the presence of large-scale geographical descriptors such as land and ocean. To better account for different statistical characteristics of variables such as temperature or wind speed, more flexible nonstationary models are needed. Jun (2014) considered nonstationary models with a differential operator approach and spatially varying smoothness parameters over land and ocean. Castruccio and Guinness (2017) also relaxed the assumption of axial symmetry by proposing an evolutionary spectrum approach to account for different regimes over land and ocean. In this work, we propose a generalization of this approach to allow spatial smoothness to change with orography, and a novel approach for changing spectral dependence across latitudes for different wavenumbers.

3.2. *The statistical framework.* Climate model variables in the atmospheric component tend to forget their initial conditions after a small number of time steps. Each ensemble member evolves in "deterministically chaotic" patterns after the climate model forgets its initial state [Lorenz (1963)]. Collins (2002), Collins and Allen (2002) and Branstator and Teng (2010) discussed the validity of the deterministically chaotic nature of climate models. Since ensemble members from the LENS differ only in their initial conditions [Kay et al. (2015)], each one will be treated as a statistical realization from a common Gaussian distribution (see Figure S2 for two normality tests for this data set). Denote by $W_r(L_m, \ell_n, t_k)$ the spatio-temporal near-surface wind speed for realization r at the latitude L_m, longitude ℓ_n and time t_k, where $r = 1, \dots, R$, $m = 1, \dots, M$, $n = 1, \dots, N$, and $k = 1, \dots, K$. Define the vector

$$\begin{aligned}\mathbf{W}_r = \{&W_r(L_1, \ell_1, t_1), \dots, W_r(L_M, \ell_1, t_1),\\ &W_r(L_1, \ell_2, t_1), \dots, W_r(L_M, \ell_N, t_K)\}^\top.\end{aligned}$$

We assume that $\mathbf{W}_r$ is independent across r conditional on its climate:

$$\mathbf{W}_r = \boldsymbol{\mu} + \boldsymbol{\varepsilon}_r, \qquad \boldsymbol{\varepsilon}_r \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta})), \tag{3.1}$$

where $\boldsymbol{\mu}$ is the space–time mean across realizations and $\boldsymbol{\theta}$ is a vector of fixed and unknown covariance parameters. By assuming independent realizations, we can estimate $\boldsymbol{\theta}$ using a restricted log-likelihood without providing any parametrization of $\boldsymbol{\mu}$. Castruccio and Stein (2013) provided the following expression for twice the negative restricted log-likelihood function:

$$\begin{aligned}2l(\boldsymbol{\theta}; \mathbf{D}) = {}& KNM(R-1)\log(2\pi) + KNM\log(R)\\ &+ (R-1)\log\left|\boldsymbol{\Sigma}(\boldsymbol{\theta})\right| + \sum_{r=1}^{R} \mathbf{D}_r^\top \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\mathbf{D}_r,\end{aligned} \tag{3.2}$$

where $\mathbf{D} = (\mathbf{D}_1^\top, \dots, \mathbf{D}_R^\top)^\top$ and $\mathbf{D}_r = \mathbf{W}_r - \overline{\mathbf{W}}$ where $\overline{\mathbf{W}} = \sum_{r=1}^{R} \mathbf{W}_r / R$. We use this expression throughout this work.

3.3. *Temporal dependence.* Let $\boldsymbol{\varepsilon}_r(t_k)$ be the vector of the stochastic component of (3.1) for time t_k and realization r. No evidence of nonstationarity in time was found, and we assume a Vector AutoRegressive of order 2 [VAR(2)] structure for $\boldsymbol{\varepsilon}_r(t_k)$, with different parameters for each spatial location. Diagnostics showed no evidence of the need for higher order autoregressive coefficients or cross-temporal dependence [Figures S3 and S4 in the supplementary material Jeong et al. (2018)]. A nonnegligible temporal dependence across locations (as observed at higher temporal resolutions) would imply a nonseparable model. Our model can be modified to allow for interactions of temporal dependence across neighboring locations [Tagle et al. (2017)]. The VAR(2) model is

$$\boldsymbol{\varepsilon}_r(t_k) = \boldsymbol{\Phi}_1\boldsymbol{\varepsilon}_r(t_{k-1}) + \boldsymbol{\Phi}_2\boldsymbol{\varepsilon}_r(t_{k-2}) + \mathbf{S}\mathbf{H}_r(t_k), \tag{3.3}$$

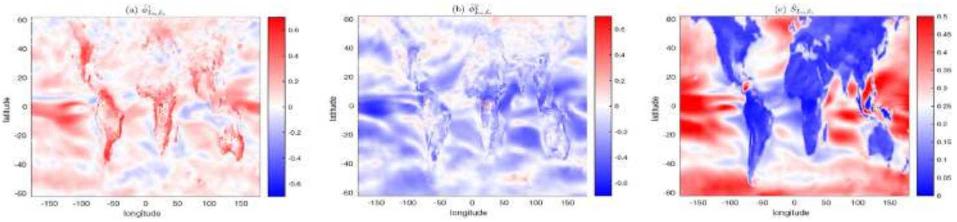

FIG. 2. *Plots of the estimated autoregressive parameters for the temporal model as defined in* (3.3): (a) $\hat{\phi}^1_{L_m,\ell_n}$, (b) $\hat{\phi}^2_{L_m,\ell_n}$ *and* (c) $\hat{S}_{L_m,\ell_n}$.

where $\mathbf{\Phi}_1 = \operatorname{diag}\{\phi^1_{L_m,\ell_n}\}$ and $\mathbf{\Phi}_2 = \operatorname{diag}\{\phi^2_{L_m,\ell_n}\}$ are two $MN \times MN$ diagonal matrices with autoregressive coefficients, and $\mathbf{S} = \operatorname{diag}\{S^1_{L_m,\ell_n}\}$ is an $MN \times MN$ diagonal matrix with the associated standard deviations, so that the temporal parameters are denoted by $\boldsymbol{\theta}_{\text{time}} = (\phi^1_{L_m,\ell_n}, \phi^2_{L_m,\ell_n}, S_{L_m,\ell_n})^\top$ for $n = 1, \dots, N$ and $m = 1, \dots, M$. For all spatial locations, we estimate $\mathbf{\Phi}_1$, $\mathbf{\Phi}_2$ and $\mathbf{S}$ by assuming that the innovations $\mathbf{H}_r(t_k) = \{H_r(L_m, \ell_n, t_k)\}$ are independent across latitude and longitude. This allows us to perform inference in parallel: each spatial location can be estimated independently by a core in a workstation or cluster. Here, $\mathbf{H}_r(t_k) \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{C})$, and the following Sections 3.4 and 3.5 are entirely devoted to determining the $\mathbf{H}_r(t_k)$ for $\boldsymbol{C}$.

Figure 2 shows the estimated autoregressive parameters. The two autoregressive coefficients, $\phi^1_{L_m,\ell_n}$ and $\phi^2_{L_m,\ell_n}$, are estimated to be mostly positive and negative, respectively [corresponding p-values are available in Figure S3 in the supplementary material Jeong et al. (2018)]. $\hat{S}_{L_m,\ell_n}$ exhibits higher values over ocean than over land. The marginal standard deviation shows similar patterns to $\hat{S}_{L_m,\ell_n}$ with a different scale (not shown).

3.4. *Longitudinal dependence.* We now provide a model for the spatial correlation of the unscaled innovations, $H_r(L_m, \ell_n, t_k)$, at different longitudes but at the same latitude. An evolutionary spectrum allows for changing behavior across large-scale geographical descriptors. Castruccio and Guinness (2017) proposed to model $H_r(L_m, \ell_n, t_k)$ in the spectral domain by performing a generalized Fourier transform across longitude:

$$H_r(L_m, \ell_n, t_k) = \sum_{c=0}^{N-1} f_{L_m,\ell_n}(c) \exp(i\ell_n c) \widetilde{H}_r(c, L_m, t_k), \tag{3.4}$$

where i is the imaginary unit, $c = 0, \dots, N-1$ is the wavenumber, $f_{L_m,\ell_n}(c)$ is the evolutionary spectrum across longitude, and $\widetilde{H}_r(c, L_m, t_k)$ is the transformed process in the spectral domain.

In this work, we propose a flexible model in which ocean, land and high mountains with altitude information are included as covariates to better account for the statistical behavior of wind speed. The United Nations Environmental Programme does not provide an unambiguous definition of “mountainous environment” [Blyth et al. (2002)]. Hence, we subjectively choose a threshold value of 1000 m [see Figure S5 in the supplementary material Jeong et al. (2018) for the global distribution of high mountains]. We allow $f_{L_m,\ell_n}(c)$ to depend on ℓ_n in a land, ocean and high mountain domain so that it can be expressed as

$$\begin{aligned}
&f_{L_m,\ell_n}(c)\\
&\qquad = f^1_{L_m,\ell_n}(c) I_{\mathrm{land}\cap\mathrm{hmt}}(L_m,\ell_n) + f^2_{L_m,\ell_n}(c) b_{\mathrm{land}\cap\mathrm{hmt}^c}(L_m,\ell_n; g_{L_m}, r_{L_m})\\
&\qquad\quad + f^3_{L_m,\ell_n}(c)\{1 - b_{\mathrm{land}}(L_m,\ell_n; g_{L_m}, r_{L_m})\},\\
&b_{\mathrm{land}}(L_m,\ell_n; g_{L_m}, r_{L_m})\\
&\qquad = \sum_{n'=1}^{N} \tilde{I}_{\mathrm{land}}(L_m,\ell_n; g_{L_m}) w(L_m, \ell_n - \ell_{n'}; r_{L_m}),
\end{aligned} \tag{3.5}$$

where $I_{\mathrm{land}\cap\mathrm{hmt}}(L_m,\ell_n)$ is the indicator function for high mountains. The transition between nonmountainous land and ocean in the second and third terms requires a parametrization for a smooth transition. Here, the modified indicator function of $I_{\mathrm{land}}(L_m,\ell_n)$ is $\tilde{I}_{\mathrm{land}}(L_m,\ell_n; g_{L_m})$, which is equal to 1 for g_{L_m} grid points wherever there is a land/ocean transition (this parameter can also be negative) and $w(L_m, \ell_n - \ell_{n'}; r_{L_m})$ is the Tukey taper function [Tukey (1967)] with range r_{L_m} (other taper functions are equally effective). Hence, $b_{\mathrm{land}}(L_m,\ell_n; g_{L_m}, r_{L_m})$ allows for a smoother transition between land/ocean states by convolving the modified land/ocean indicator, $\tilde{I}_{\mathrm{land}}(L_m,\ell_n; g_{L_m})$, with the taper function, $w(L_m, \ell_n - \ell_{n'}; r_{L_m})$. We additionally use the information of the surface altitude, which has an impact on land and high mountains. The component spectra in (3.5) is defined according to the parametric form [Castruccio and Stein (2013), Poppick and Stein (2014)]:

$$|f^j_{L_m,\ell_n}(c)|^2 = \phi^j_{L_m,\ell_n}\{(\alpha^j_{L_m,\ell_n})^2 + 4\sin^2(c\pi/N)\}^{\nu^j_{L_m,\ell_n}+1/2}, \qquad j = 1, 2, 3,$$

where $(\phi^j_{L_m,\ell_n}, \alpha^j_{L_m,\ell_n}, \nu^j_{L_m,\ell_n})$ have a similar interpretation as the variance, inverse range, and smoothness parameters, respectively, for the Matérn spectrum over the line. We allow spatially varying parameters to depend on the surface altitude, with log-linear parametrization to ensure positivity for $\phi^j_{L_m,\ell_n} = \beta^{j,\phi}_{L_m} \exp[\tan^{-1}\{A_{L_m,\ell_n}\gamma^{\phi}_{L_m}\}]$, $j = 1, 2$ and $\phi^3_{L_m,\ell_n} = \beta^{3,\phi}_{L_m}$, where $\beta^{j,\phi}_{L_m}$ is a positive number, $\gamma^{\phi}_{L_m}$ is a real number and A_{L_m,ℓ_n} represents the altitude at location (L_m, ℓ_n). $\nu^j_{L_m,\ell_n}$ and $\alpha^j_{L_m,\ell_n}$ have a similar structure. In order to avoid overparametrization, $\gamma^{\phi}_{L_m}$ controls the impact of the surface altitude for land and high mountains, that is, $\phi^1_{L_m,\ell_n}(c)$ and $\phi^2_{L_m,\ell_n}(c)$ share the same coefficient, $\gamma^{\phi}_{L_m}$. Hence, the longitudinal parameters are $\boldsymbol{\theta}_{\mathrm{lon}} = (\beta^{j,\phi}_{L_m}, \gamma^{\phi}_{L_m}, \beta^{j,\nu}_{L_m}, \gamma^{\nu}_{L_m}, \beta^{j,\alpha}_{L_m}, \gamma^{\alpha}_{L_m}, g_{L_m},$

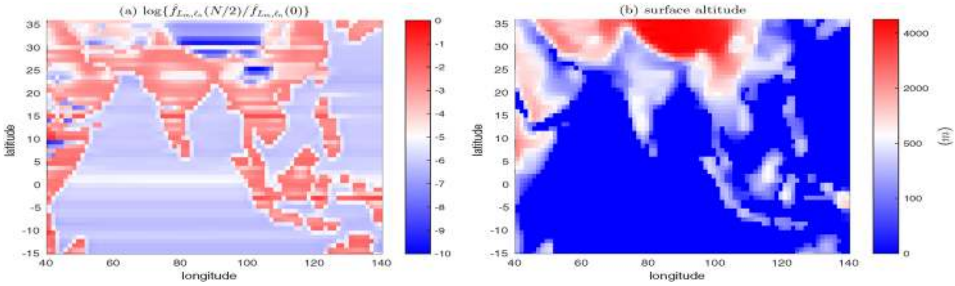

FIG. 3. (a) *Log-ratio of periodograms*, $\log\{\hat{f}_{L_m,\ell_n}(N/2)/\hat{f}_{L_m,\ell_n}(0)\}$, *and* (b) *surface altitude* (*orography*) *near the Indian Ocean and Himalayan region*.

$r_{L_m})^\top$, $j = 1, 2, 3$ and $m = 1, \ldots, M$. The parameter values for each L_m are independent from the other latitudinal bands, therefore, each core of a workstation or cluster can perform inference independently on each band.

In Figure 3(a), we show $\log\{\hat{f}_{L_m,\ell_n}(N/2)/\hat{f}_{L_m,\ell_n}(0)\}$, the log-ratio of periodograms that empirically estimates the rate of spectral decay at high frequency, and the surface altitude near the Indian Ocean and Himalayan region. At high altitudes, the Himalayan region and Western China exhibit pronounced spectral decay compared to neighboring land masses at low altitudes, such as India and Eastern China. Moreover, the patterns of spectral decay markedly follow the topographical relief, as apparent from Figure 3(b). Indeed, besides a smoother ocean behavior, annual winds are considerably smoother at high altitudes, as demonstrated by the fast rate of spectral decay over the region corresponding to the Himalayas.

Figure 4 presents a comparison of three models: the axially symmetric model (AX), the evolutionary spectrum model with land and ocean (LAO) and the new evolutionary spectrum model with altitude (ALT), in terms of the Bayesian Infor-

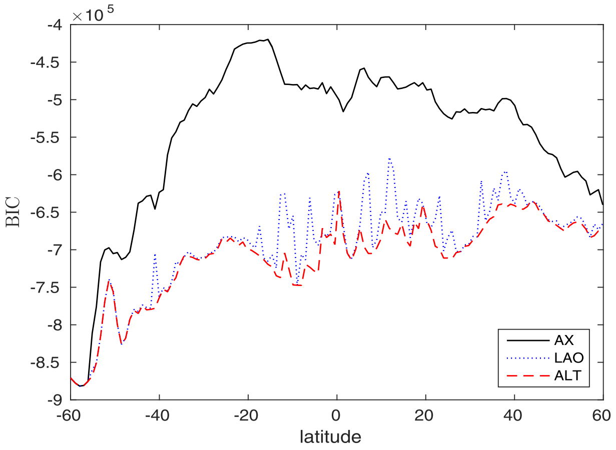

FIG. 4. *Comparison of AX, LAO and ALT models in terms of BIC versus latitude.*

mation Criterion (BIC) against latitude. LAO and ALT uniformly outperform AX, but ALT is significantly more flexible than LAO at latitudinal bands between 25°S and 45°N, where the percentage of points with high mountains within these bands is 7.6%, compared to 3% within the other bands.

3.5. *Latitudinal dependence.* We propose a novel Vector AutoRegressive model of order 1, VAR(1), across latitudes to allow for dependence of $\widetilde{H}_r(c, L_m, t_k)$ across neighboring wavenumbers. For any r and t_k, denote by $\widetilde{\mathbf{H}}_{L_m} = \{\widetilde{H}_{L_m}(c_1), \dots, \widetilde{H}_{L_m}(c_N)\}^\top$, then

$$\widetilde{\mathbf{H}}_{L_m} = \begin{cases} \boldsymbol{\varphi}_{L_m} \widetilde{\mathbf{H}}_{L_{m-1}} + \mathbf{e}_{L_m}, & m = 2, \dots, M, \\ \mathbf{e}_{L_1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), & m = 1, \end{cases} \tag{3.6}$$

$$\mathbf{e}_{L_m} \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{L_m}), \qquad m > 1,$$

where $\boldsymbol{\varphi}_{L_m}$ is an $N \times N$ matrix describing the autoregressive coefficients and $\boldsymbol{\Sigma}_{L_m}$ in an $N \times N$ matrix with the covariance structure of the innovation. We propose the following banded structure, which eases the computational burden by inducing sparsity and also results in a diagonally dominant matrix:

$$\boldsymbol{\varphi}_{L_m} = \begin{pmatrix} \varphi_{L_m}(c_1) & \frac{\{1-\varphi_{L_m}(c_1)\}a_{L_m}}{4} & \frac{\{1-\varphi_{L_m}(c_1)\}b_{L_m}}{4} & 0 & \cdots & 0 & 0 & 0 \\ \frac{\{1-\varphi_{L_m}(c_2)\}a_{L_m}}{4} & \varphi_{L_m}(c_2) & \frac{\{1-\varphi_{L_m}(c_2)\}a_{L_m}}{4} & \frac{\{1-\varphi_{L_m}(c_2)\}b_{L_m}}{4} & \cdots & 0 & 0 & 0 \\ \frac{\{1-\varphi_{L_m}(c_3)\}b_{L_m}}{4} & \frac{\{1-\varphi_{L_m}(c_3)\}a_{L_m}}{4} & \varphi_{L_m}(c_3) & \frac{\{1-\varphi_{L_m}(c_3)\}a_{L_m}}{4} & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \frac{\{1-\varphi_{L_m}(c_{N-1})\}a_{L_m}}{4} & \varphi_{L_m}(c_{N-1}) & \frac{\{1-\varphi_{L_m}(c_{N-1})\}a_{L_m}}{4} \\ 0 & 0 & 0 & 0 & \cdots & \frac{\{1-\varphi_{L_m}(c_N)\}b_{L_m}}{4} & \frac{\{1-\varphi_{L_m}(c_N)\}a_{L_m}}{4} & \varphi_{L_m}(c_N) \end{pmatrix}, \tag{3.7}$$

where $a_{L_m}, b_{L_m} \in (-1, 1)$ for all m, $\boldsymbol{\Sigma}_{L_m} = \operatorname{diag}\{1 - \varphi_{L_m}(c_n)^2\}$ and

$$\varphi_{L_m}(c) = \frac{\xi_{L_m}}{\{1 + 4\sin^2(c\pi/N)\}^{\tau_{L_m}}}, \tag{3.8}$$

where $\xi_{L_m} \in [0, 1]$ and $\tau_{L_m} > 0$ for all m. If $a_{L_m} = b_{L_m} = 0$, this model corresponds to a nonstationary AR(1) process in latitude:

$$\operatorname{corr}\{\widetilde{H}_r(c, L_m, t_k), \widetilde{H}_{r'}(c', L_{m'}, t_{k'})\} = \mathbf{1}\{c = c', k = k', r = r'\}\rho_{L_m, L_{m'}}(c),$$

where $\rho_{L_m, L_{m'}}(c) = \prod_{j=m}^{m'} \varphi_{L_j}(c)$, $m < m'$ is the coherence between latitudes L_m and $L_{m'}$ among the $\widetilde{H}_r(c, L_m, t_k)$s with the same wavenumber, time and realization [Castruccio and Guinness (2017)].

To compare VAR(1) with AR(1), we perform inference for every pair of contiguous bands (L_m, L_{m+1}) independently for both models, and we report the BIC and parameter estimates in Figures 5(a) and (b), respectively. For most latitudes, VAR(1) has a large BIC improvement compared with AR(1), and $\hat{a}_{L_m}$ and $\hat{b}_{L_m}$ are significantly different from 0 (see confidence bands).

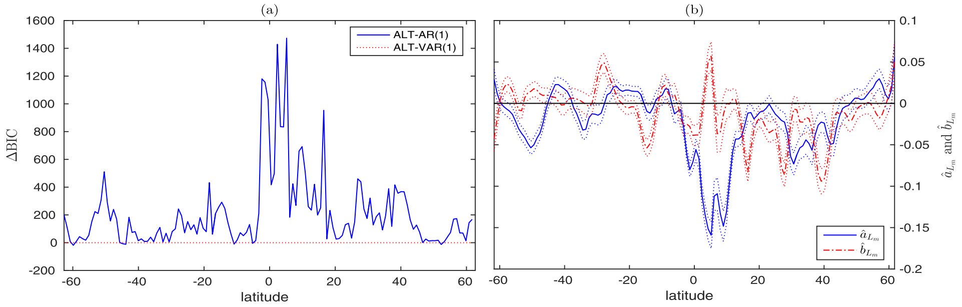

FIG. 5. *Comparison between* AR(1) *and* VAR(1) *latitudinal models for adjacent bands in terms of* (a) *BIC and* (b) $\hat{a}_{L_m}$ *and* $\hat{b}_{L_m}$ *as in* (3.7) (*the dotted lines represent the* 95% *confidence bands*). *A smoothing spline has been applied to the parameters estimated in* (b).

To complete the model, the latitudinal dependence of a_{L_m}, b_{L_m} in (3.7) and ξ_{L_m}, τ_{L_m} in (3.8) must be specified. Figure 5(b) highlights how latitudes near the equator result in $\hat{a}_{L_m}$ and $\hat{b}_{L_m}$ being considerably (and significantly) different from zero, hence the need of different coefficients near these latitudinal bands. To mitigate, however, the increased computational cost derived from these additional parameters we choose the bounds $-30°$ and $30°$, consistently with Castruccio and Guinness (2017), in order to include the tropics, whose climate is determined by the complex interactions between large-scale atmospheric circulation, atmospheric convection, solar and terrestrial radiactive transfer, boundary layers and clouds [Betts and Ridgway (1988)]. As an important indicator of atmospheric circulation, wind in these bands is influenced by the Hadley and Walker circulations, which are the mean meridional and longitudinal overturning circulations, respectively. In particular, the Walker circulation is affected by the El Niño-Southern Oscillation (ENSO) over the Pacific Ocean [Gastineau, Li and Le Treut (2009)]. Therefore, for $-30° < L_m < 30°$ we assume that (ξ_{L_m}, τ_{L_m}) are fixed and equal to the estimated value from the adjacent band fit in Figure 5, whereas we assume a constant value equal to (ξ, τ) outside this range and (a,b) for all latitudinal bands. The parameter estimates and corresponding 95% confidence intervals are $\hat{a} = 0.136$ $(0.132, 0.140)$, $\hat{b} = 0.071$ $(0.067, 0.075)$, $\hat{\xi} = 0.960$ $(0.903, 1.000)$ and $\hat{\tau} = 0.628$ $(0.626, 0.630)$. The latitudinal parameters are then $\boldsymbol{\theta}_{\text{lat}} = (a, b, \xi_{L_m}, \tau_{L_m})^\top$ for m such that the latitudes are in the range of $-30° < L_m < 30°$. They are otherwise constant.

3.6. *Inference.* A computational benefit of axially symmetric models on regularly spaced data is that the resulting covariance matrix is block circulant, and hence block diagonal in the spectral domain [Davis (1979)]. Thus, likelihood evaluation is convenient in the spectral domain, requiring matrix inversion and determinant computation of small matrices [Jun and Stein (2008)]. In case of a nonstationary model across longitude at a given latitude, it is still possible to derive a likelihood expression whose computational efficiency is close to that of the axially symmetric case if the data are on a regular grid.

Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_{\mathrm{time}}^\top, \boldsymbol{\theta}_{\mathrm{lon}}^\top, \boldsymbol{\theta}_{\mathrm{lat}}^\top)^\top$, where $\boldsymbol{\theta}_{\mathrm{time}}$, $\boldsymbol{\theta}_{\mathrm{lon}}$, and $\boldsymbol{\theta}_{\mathrm{lat}}$ are collections of all temporal, longitudinal and latitudinal parameters, respectively. If the data are on a grid, (3.2) simplifies to

$$\begin{aligned} 2l(\boldsymbol{\theta}; \mathbf{D}) &= KNM(R-1)\log(2\pi) + KNM\log(R) \\ &\quad + (R-1)\sum_{m=1}^{M} \log\left|\boldsymbol{\Sigma}_m^1(\boldsymbol{\theta}_{\mathrm{lon}})\right| + (R-1)\sum_{p=1}^{P} \log\left|\boldsymbol{\Sigma}_p^2(\boldsymbol{\theta}_{\mathrm{lat}})\right| \\ &\quad + \sum_{r=1}^{R}\sum_{k=1}^{K}\sum_{p=1}^{P} \mathbf{v}_p(t_k, r; \boldsymbol{\theta}_{\mathrm{time}}, \boldsymbol{\theta}_{\mathrm{lon}})^\top \boldsymbol{\Sigma}_p^2(\boldsymbol{\theta}_{\mathrm{lat}})^{-1} \mathbf{v}_p(t_k, r; \boldsymbol{\theta}_{\mathrm{time}}, \boldsymbol{\theta}_{\mathrm{lon}}), \end{aligned} \tag{3.9}$$

where $\boldsymbol{\Sigma}_m^1(\boldsymbol{\theta}_{\mathrm{lon}})$ is the $N \times N$ coherence matrix of latitudinal band L_m, $\boldsymbol{\Sigma}_p^2(\boldsymbol{\theta}_{\mathrm{lat}})$ is the $(M \times \lfloor N/P \rfloor) \times (M \times \lfloor N/P \rfloor)$ covariance matrix describing the coherence among multiple latitudinal bands, which is obtained by approximating $\boldsymbol{\varphi}_{L_m}$ in (3.7) with $p = 1, \dots, P$ diagonal blocks and the vector $\mathbf{v}_p(t_k, r; \boldsymbol{\theta}_{\mathrm{time}}, \boldsymbol{\theta}_{\mathrm{lon}})$ is a suitable transformation of $\mathbf{D}$ [Castruccio and Genton (2014)]. To estimate the spatial and temporal structure of the data, we use (3.9) throughout this study.

As $\boldsymbol{\theta}$ is typically very high dimensional, we achieve an approximate maximum likelihood estimator by applying (3.9) under a conditional approximations inference scheme that assumes independence across increasingly large subsets, as in Castruccio and Stein (2013). Each approximation assumes that the parameters obtained from previous steps are fixed and known for the upcoming steps:

Step 1. Estimate the temporal parameters, $\boldsymbol{\theta}_{\mathrm{time}}$, by assuming that there is no cross-temporal dependence in latitude and longitude;

Step 2. Consider that $\boldsymbol{\theta}_{\mathrm{time}}$ is fixed at its estimated value and estimate $\boldsymbol{\theta}_{\mathrm{lon}}$ by assuming that the latitudinal bands are independent;

Step 3. Consider $\boldsymbol{\theta}_{\mathrm{time}}$ and $\boldsymbol{\theta}_{\mathrm{lon}}$ fixed at their estimated values and estimate $\boldsymbol{\theta}_{\mathrm{lat}}$.

Since steps 1 and 2 assume independence across subsets, inference can be performed independently by multiple processors in a workstation or in a cluster.

As argued by Castruccio and Guinness (2017), the sequential approach with previously estimated parameters could produce an estimation bias. This is mostly apparent from step 2 to 3, where the estimated parameters for the single latitudinal band approximation may not be the optimal values for the multiple latitudinal band approximation. One solution to mitigate this issue is to refit $\boldsymbol{\theta}_{\mathrm{lon}}$ for two adjacent bands. This step requires additional computational time, 1.5 to 2 hours on a 24-cores workstation for the ALT-VAR model (parallelizing the inference for different sets of contiguous bands) but it improved model fit markedly in this study. This can be done for several adjacent bands if the computational time is acceptable, but refitting all bands with the full data set may require several weeks of computational time and very powerful computational resources.

4. Model comparison and validation of local behavior. We compare the model introduced in the previous section with previously available models, and we validate the local space–time structure against the data.

Table 1 presents a comparison in terms of model selection metrics: a land/ocean evolutionary spectrum with a nonstationary latitudinal AR(1) process [LAO-AR(1)], our new evolutionary spectrum with a nonstationary latitudinal AR(1) process [ALT-AR(1)] and with a nonstationary latitudinal VAR(1) process [ALT-VAR(1)]. ALT-AR(1) requires approximately 1.67 times more parameters than does LAO-AR(1), but it shows clear improvements in terms of the normalized log-likelihood, BIC and other standard model selection metrics (not shown). ALT-AR(1) allows for spatially varying coefficients across the mountain profiles and shows a noticeable improvement in model fit as the log-likelihood improves by 0.08 units per observation. The most general ALT-VAR(1) requires two additional parameters a and b, and it achieves a further improvement in the fit. While the relative improvement between ALT-VAR(1) and ALT-AR(1) compared to the improvement between LAO-AR(1) and ALT-AR(1) is not conspicuous, the results in Table 1 are expressed in 10^8 units and, as Figure 5(a) highlights, the improvement in absolute terms is far from being negligible: the BIC improves hundreds, or even thousands of units in some latitudes.

We assess the high-frequency behavior of the models by computing the contrast variances to assess the quality of the fit [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 compute the squared distances between the empirical and fitted variances for both LAO-AR(1) and ALT-VAR(1), and plot their differences in Figure 6. Positive

TABLE 1

Comparison between different models in terms of the number of parameters (*excluding the temporal component*), *the normalized restricted log-likelihood*, *and BIC. The general guidelines for* Δloglik$/\{NMK(R-1)\}$ *are that anything above* 0.1 *is large and anything above* 0.01 *is modest but still sizable* [*Castruccio and Stein* (2013)]

Model	**LAO-AR(1)**	**ALT-AR(1)**	**ALT-VAR(1)**
# of parameters	1202	2006	2008
Δloglik$/\{NMK(R-1)\}$	0	0.08152	**0.08177**
BIC ($\times 10^8$)	−1.02638	−1.05015	**−1.05023**

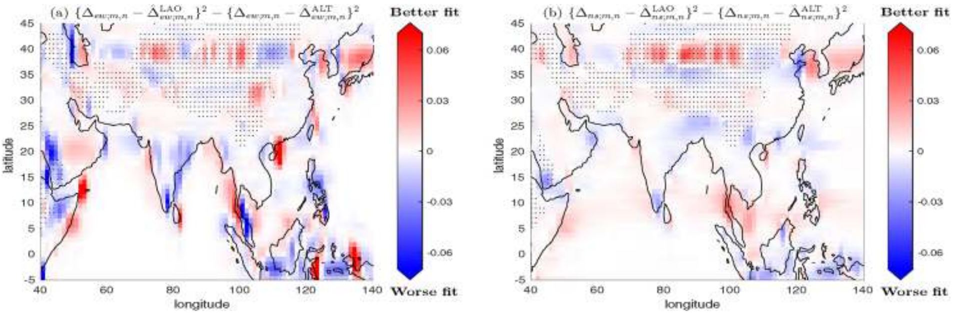
(a) $\{\Delta_{ew;m,n} - \hat{\Delta}^{\text{LAO}}_{ew;m,n}\}^2 - \{\Delta_{ew;m,n} - \hat{\Delta}^{\text{ALT}}_{ew;m,n}\}^2$
(b) $\{\Delta_{ns;m,n} - \hat{\Delta}^{\text{LAO}}_{ns;m,n}\}^2 - \{\Delta_{ns;m,n} - \hat{\Delta}^{\text{ALT}}_{ns;m,n}\}^2$
Better fit
Worse fit
0.06
0.03
0
-0.03
-0.06
latitude
longitude
45
40
35
30
25
20
15
10
5
0
-5
40
60
80
100
120
140

FIG. 6. *The squared distances of the fitted contrast variances from the empirical contrast variances between two models*, LAO-AR(1) *and* ALT-VAR(1): (a) $\{\Delta_{ew;m,n} - \hat{\Delta}^{\text{LAO}}_{ew;m,n}\}^2 - \{\Delta_{ew;m,n} - \hat{\Delta}^{\text{ALT}}_{ew;m,n}\}^2$ *and* (b) $\{\Delta_{ns;m,n} - \hat{\Delta}^{\text{LAO}}_{ns;m,n}\}^2 - \{\Delta_{ns;m,n} - \hat{\Delta}^{\text{ALT}}_{ns;m,n}\}^2$. *Black dots indicate the locations where the surface altitude is larger than* 1000 *m*.

and negative values represent better and worse model fit of ALT-VAR(1) compared to LAO-AR(1), respectively. The Himalayan region (from 78.75°E to 86.25°E and from 26.86°N to 30.63°N) has considerably more positive values for the north-south contrast variance case in Figure 6(b). It is also apparent how ALT-VAR(1) shows a better model fit near the Tian Shan mountain region (from 72.5°E to 80°E and from 38.16°N to 41°N) with positive values for both east-west and north-south contrast variance cases.

To quantify the improvement corresponding to these mountain ranges, we computed the aforementioned difference among these two mountain regions and compared their distributions. Table 2 represents the 25th, 50th, 75th percentiles of difference near Himalayan and Tian Shan mountain regions, and we observe that overall the distributions tend to have more positive values, that is, ALT-VAR(1) has better model fit in terms of contrast variances compared to LAO-AR(1). The table also confirms the visual inspection in Figure 6: the two metrics have larger values near Tian Shan mountain region compared to near Himalayan region.

TABLE 2
25th, 50th and 75th percentiles of two difference metrics near Himalayan region (H) and Tian Shan mountain region (T)

Metric	**Region**	**25th**	**50th**	**75th**
$[\{\Delta_{ew;m,n} - \hat{\Delta}^{\text{LAO}}_{ew;m,n}\}^2 - \{\Delta_{ew;m,n} - \hat{\Delta}^{\text{ALT}}_{ew;m,n}\}^2] \times 10^3$	H	−1	1	2
	T	−9	20	57
$[\{\Delta_{ns;m,n} - \hat{\Delta}^{\text{LAO}}_{ns;m,n}\}^2 - \{\Delta_{ns;m,n} - \hat{\Delta}^{\text{ALT}}_{ns;m,n}\}^2] \times 10^3$	H	0	6	10
	T	−3	8	52

5. Generation of stochastic surrogates and validation of large-scale behavior. In this section, we explain how to generate the stochastic surrogates from the SG. Besides their interest for wind energy assessment, such surrogate runs can then be compared with the original LENS runs to validate the large-scale behavior of the statistical model.

In the previous sections, $\boldsymbol{\theta} = (\boldsymbol{\theta}_{\text{time}}^\top, \boldsymbol{\theta}_{\text{lon}}^\top, \boldsymbol{\theta}_{\text{lat}}^\top)^\top$ in (3.1) have been defined and estimated from the training set. The mean climate $\boldsymbol{\mu}$ can be obtained as a smoothed version of the ensemble mean $\overline{\mathbf{W}}$. Similar to Castruccio and Genton (2016) and Castruccio and Guinness (2017), for each location (L_m, ℓ_n) we fit a smoothing spline $\widetilde{W}(L_m, \ell_n, t_k)$ for $k = 1, \dots, K$, which minimizes

$$\lambda \sum_{k=1}^{K} \{\overline{W}(L_m, \ell_n, t_k) - \widetilde{W}(L_m, \ell_n, t_k)\}^2 + (1 - \lambda) \sum_{k=1}^{K} \{\nabla_2 \widetilde{W}(L_m, \ell_n, t_k)\}^2,$$

where ∇_2 is the second-order finite difference operator. We impose a penalty term, $\lambda = 0.01$, to reflect the slowly varying climate of annual wind fields over the next century [Vaughan and Cracknell (2013)].

Once $\boldsymbol{\mu}$ and $\boldsymbol{\theta}$ are estimated, surrogate runs can be almost instantaneously generated on a modest laptop by performing the following steps:

Step 1. Generate $\mathbf{e}_{L_m} \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{L_m})$ as in (3.6);
Step 2. Compute $\widetilde{\mathbf{H}}_{L_m}$ with expressions (3.6);
Step 3. Compute $H_r(L_m, \ell_n, t_k)$ with expression (3.4);
Step 4. Compute $\boldsymbol{\varepsilon}_r$ with equation (3.3);
Step 5. Obtain the reproduced run as $\widetilde{\mathbf{W}} + \boldsymbol{\varepsilon}_r$, where

$$\widetilde{\mathbf{W}} = \{\widetilde{W}(L_1, \ell_1, t_1), \dots, \widetilde{W}(L_M, \ell_1, t_1), \widetilde{W}(L_1, \ell_2, t_1), \dots, \widetilde{W}(L_M, \ell_N, t_K)\}^\top.$$

We generated one hundred runs and compared them with the climate model runs; see Figure S8 for a comparison in 2050 of five runs with other five LENS runs not in the training set and a movie of a surrogate run (Movie S1). We computed near-future (2013–2046) annual wind speed trends (a reference metric in the reference LENS publication [Kay et al. (2015)]) for each of the surrogate and LENS runs and then plotted the corresponding means in Figures 7(a) and 7(b) (see Figure S7 for a comparison of the individual runs), and the 2.5th, 50th and 97.5th percentiles in 2050 in Figure S9. From these figures, it is apparent how the SG and LENS distributions are visually indistinguishable, with a stronger trend over ocean and coastline than over land.

Figure 7(c)–(d) shows a comparison between reproduced and climate model runs in terms of their distribution of wind power density at 80 m in 2020 [details on how to derive this variable from wind speed are provided in the supplementary Jeong et al. (2018)] for locations near Riyadh (24.97°N and 46.25°E) and Rabigh (23.01°N and 38.75°E), Saudi Arabia. Both locations are in the Arabian peninsula and exhibit significant nondecreasing trends. So, an assessment of the internal

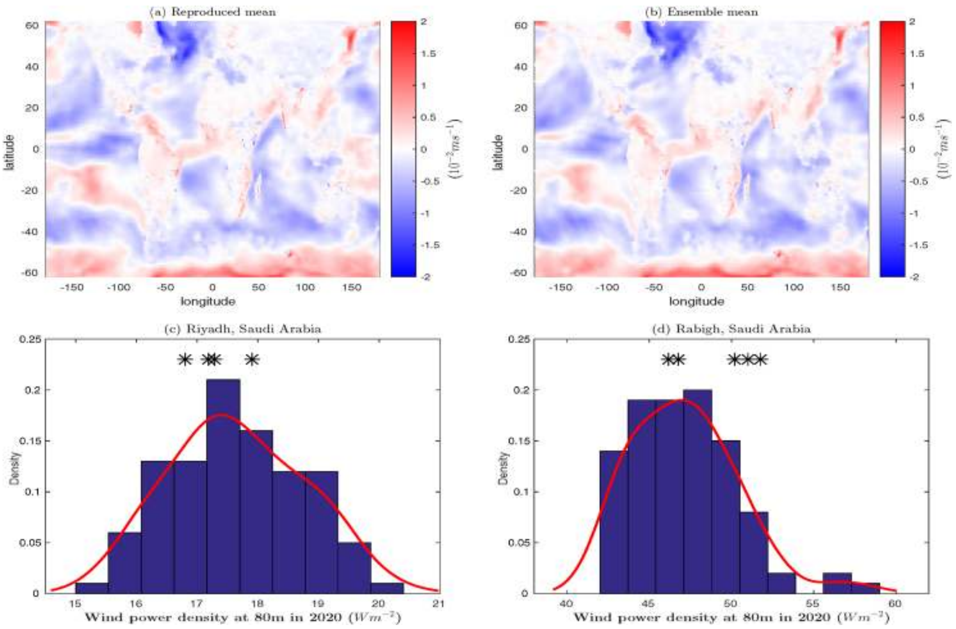

FIG. 7. *Top*: *Global maps of* (a) *the mean from reproduced runs and* (b) *the ensemble mean of the near-future* (2013–2046) *annual near-surface wind speed trends*. *Bottom*: *Histogram of the distribution of the wind power density at* 80 *m in* 2020 *with nonparametric density in red for the one-hundred reproduced runs near* (c) *Riyadh and* (d) *Rabigh*, *Saudi Arabia* ($*$ *represents the original climate model runs*).

variability is crucial to determining the robustness of the point estimates and could inform policy makers on the uncertainty and associated risks in building wind turbines in these areas where no regional studies and very limited ground-based observations are available. Here, we observe that Rabigh on the coastline has considerably more potential to generate wind power than Riyadh in the central inland of Saudi Arabia. A more accurate assessment of wind resources could be achieved by using wind speed data at a higher spatio-temporal resolution than the one used in this study (i.e., annual mean wind speed at horizontal resolution of approximately 1°), but such an assessment is currently unfeasible given the absence of ESM simulations at fine spatio-temporal resolutions for multiple decades. The five climate model runs are poorly informative for internal variability, but the distribution generated from many reproduced runs allows for a more accurate assessment. Both locations exhibit a considerable variability in wind power density (2.5 and 97.5 percentiles), with (15.7, 19.7) Wm^{-2} for Riyadh and (42.3, 55.9) Wm^{-2} for Rabigh.

Figure 7(c)–(d) depends only on the marginal wind at two given locations, so it could be obtained with simpler pointwise approaches without assuming spatial dependence. The SG, however, allows to generate spatially resolved fields, which are indistinguishable from the original LENS runs (see Figures S7 and S8). To visualize this interactively, a dynamic Graphical User Interface (GUI) application in Matlab is provided in the supplementary material [Jeong et al. (2018)]. The GUI requires to download $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\theta}}$ in (3.1), for a total of 30 megabytes, instead of downloading the entire climate model ensemble (40 members), which is 1.1 gigabytes. A user can then use the stored coefficients and generate many runs to achieve a considerably more detailed assessment of wind uncertainty under different initial conditions.

6. Discussion and conclusion. Understanding the spatio-temporal variability of wind resources is essential to sustain the increasing energy demand, but traditional ESM ensemble-based approaches for assessment in developing countries are increasingly computationally, time and memory consuming. SGs provide a simple and computationally convenient tool for generating surrogate runs under different initial conditions and assessing the uncertainty from internal variability without storing a prohibitive amount of information. Once inference is performed and the parameters have been estimated from a small number of LENS members, an end user can download a small software package and use it to almost instantaneously generate many reproduced runs whose large-scale features are almost identical to the original runs [see Figures 7(a) and (b)] and assess the uncertainty in future wind power density due to internal variability [see Figure 7(c) and (d)].

We introduced a spectral model for gridded data which allows for an improved fit of global wind data. Our proposed model presents two elements of novelty from the current literature:

1. It incorporates more large-scale geographical information to explain the nonstationary behavior of wind across longitude. In particular, the model incorporates orography, which is shown to affect the spatial smoothness of wind fields. The proposed model allows for spatially varying parameters depending on the surface altitude over land and high mountains, contains the axially symmetric and the land/ocean evolutionary spectrum as special cases and shows improved performance in terms of the log-likelihood, BIC and other standard model selection metrics.
2. It introduces a nonstationary VAR(1) model for the latitudinal coherence for multiple wavenumbers. By assuming independent partitions of the correlated innovations for neighboring wavenumbers, the proposed model still holds a convenient formulation of the log-likelihood function in (3.9) and further improves the model fit.

Inference is performed via a multi-step conditional likelihood approach, which leverages on parallel computation and achieves a fit on a data set of more than 18 million data points.

For policy making purposes, a clear limitation of our approach is the coarse time scale at which wind power density is assessed. Finer time scales require considerable modeling and face computational challenges. On the modeling side, the Gaussianity assumption has to be relaxed at higher temporal resolution and requires alternative trans-Gaussian processes, such as Tukey g-and-h random fields [Xu and Genton (2017)]. On the computational side, the already considerable data size of this application (more than 18 million data points) will be increased by more than two orders of magnitude. While clearly adding a layer of complexity to inference, the same key ingredients, namely leveraging on regular geometries, parallel computing and spectral methods have already shown to achieve inference from data sets larger than one billion data points [Castruccio and Genton (2016)], so a global inference of daily wind power density for the entire ensemble is likely achievable with current computational architectures. If a smaller region such as Saudi Arabia is chosen, then the decrease in the number of spatial locations alleviates the computational burden to some extent, and would allow to model non-Gaussian processes at finer scale; see Tagle et al. (2017).

SUPPLEMENTARY MATERIAL

Supplement to "Reducing storage of global wind ensembles with stochastic generators" (DOI: 10.1214/17-AOAS1105SUPP; .zip). Further technical details and a Graphical User Interface application in Matlab can be found in the online supplementary material.

REFERENCES

BAKER, A. H., XU, H., DENNIS, J. M., LEVY, M. N., NYCHKA, D., MICKELSON, S. A., EDWARDS, J., VERTENSTEIN, M. and WEGENER, A. (2014). A methodology for evaluating the impact of data compression on climate simulation data. In *Proceedings of the* 23*rd International Symposium on High-Performance Parallel and Distributed Computing HPDC* '14 203–214. ACM.

BAKER, A. H., HAMMERLING, D. M., MICKELSON, S. A., XU, H., STOLPE, M. B., NAVEAU, P., SANDERSON, B., EBERT-UPHOFF, I., SAMARASINGHE, S., DE SIMONE, F., CARBONE, F., GENCARELLI, C. N., DENNIS, J. M., KAY, J. E. and LINDSTROM, P. (2016). Evaluating lossy data compression on climate simulation data within a large ensemble. *Geosci. Model Dev.* **9** 4381–4403.

BANUELOS-RUEDAS, F., ANGELES-CAMACHO, C. and RIOS-MARCUELLO, S. (2011). *Methodologies Used in the Extrapolation of Wind Speed Data at Different Heights and Its Impact in the Wind Energy Resource Assessment in a Region*. INTECH Open Access Publisher.

BARTHELMIE, R. J. and PRYOR, S. C. (2014). Potential contribution of wind energy to climate change mitigation. *Nature Climate Change* **4** 684–688.

BETTS, A. K. and RIDGWAY, W. (1988). Coupling of the radiative, convective, and surface fluxes over the equatorial pacific. *J. Atmos. Sci.* **45** 522–536.

BLYTH, S., GROOMBRIDGE, B., LYSENKO, I., MILES, L. and NEWTON, A. (2002). *Mountain Watch*: *Environmental Change & Sustainable Developmental in Mountains*. UNEP-WCMC, Cambridge.

BOLIN, D. and LINDGREN, F. (2011). Spatial models generated by nested stochastic partial differential equations, with an application to global ozone mapping. *Ann. Appl. Stat.* **5** 523–550.

BRANSTATOR, G. and TENG, H. (2010). Two limits of initial-value decadal predictability in a CGCM. *J. Climate* **23** 6292–6311.

CASTRUCCIO, S. and GENTON, M. G. (2014). Beyond axial symmetry: An improved class of models for global data. *Stat* **3** 48–55.

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 GUINNESS, J. (2017). An evolutionary spectrum approach to incorporate large-scale geographical descriptors on global processes. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **66** 329–344. MR3611690

CASTRUCCIO, S. and STEIN, M. L. (2013). Global space–time models for climate ensembles. *Ann. Appl. Stat.* **7** 1593–1611. MR3127960

CLARKE, J., ALEGRÍA, A. and PORCU, E. (2016). Regularity properties and simulations of Gaussian random fields on the sphere cross time. ArXiv Preprint arXiv:1611.02851.

COLLINS, M. (2002). Climate predictability on interannual to decadal time scales: The initial value problem. *Clim. Dyn.* **19** 671–692.

COLLINS, M. and ALLEN, M. R. (2002). Assessing the relative roles of initial and boundary conditions in interannual to decadal climate predictability. *J. Climate* **15** 3104–3109.

DAS, B. (2000). Global covariance modeling: A deformation approach to anisotropy. Ph.D. thesis, Univ. Washington.

DAVIS, P. J. (1979). *Circulant Matrices*. Wiley, New York. MR0543191

GASTINEAU, G., LI, L. and LE TREUT, H. (2009). The Hadley and Walker circulation changes in global warming conditions described by idealized atmospheric simulations. *J. Climate* **22** 3993–4013.

GNEITING, T. (2013). Strictly and non-strictly positive definite functions on spheres. *Bernoulli* **19** 1327–1349.

GUINNESS, J. and HAMMERLING, D. (2017). Compression and conditional emulation of climate model output. *J. Amer. Statist. Assoc.* To appear.

HEATON, M. J., KATZFUSS, M., BERRETT, C. and NYCHKA, D. W. (2014). Constructing valid spatial processes on the sphere using kernel convolutions. *Environmetrics* **25** 2–15.

HITCZENKO, M. and STEIN, M. L. (2012). Some theory for anisotropic processes on the sphere. *Stat. Methodol.* **9** 211–227.

JEONG, J., JUN, M. and GENTON, M. G. (2017). Spherical process models for global spatial statistics. *Statist. Sci.* **32** 501–513. MR3730519

JEONG, J., CASTRUCCIO, S., CRIPPA, P. and GENTON, M. G. (2018). Supplement to "Reducing storage of global wind ensembles with stochastic generators." DOI:10.1214/17-AOAS1105SUPP.

JONES, R. H. (1963). Stochastic processes on a sphere. *Ann. Math. Stat.* **34** 213–218.

JUN, M. (2011). Non-stationary cross-covariance models for multivariate processes on a globe. *Scand. J. Stat.* **38** 726–747.

JUN, M. (2014). Matérn-based nonstationary cross-covariance models for global processes. *J. Multivariate Anal.* **128** 134–146.

JUN, M. and STEIN, M. L. (2007). An approach to producing space–time covariance functions on spheres. *Technometrics* **49** 468–479. MR2394558

JUN, M. and STEIN, M. L. (2008). Nonstationary covariance models for global data. *Ann. Appl. Stat.* **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. *Bull. Am. Meteorol. Soc.* **96** 1333–1349.

LINDGREN, F., RUE, H. and LINDSTRÖM, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 423–498.

LORENZ, E. N. (1963). Deterministic nonperiodic flow. *J. Atmos. Sci.* **20** 130–141.

MA, C. (2015). Isotropic covariance matrix functions on all spheres. *Math. Geosci.* **47** 699–717.

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. *Atmos. Sci. Lett.* **12** 325–333.

MEARNS, L., HULME, M., CARTER, T., LEEMANS, R., LAL, M. and WHETTON, P. (2001). Climate scenario development. In *Climate Change* 2001: *The Scientific Basis*. Cambridge Univ. Press, Cambridge.

MOOMAW, W., YAMBA, F., KAMIMOTO, M., MAURICE, L., NYBOER, J., URAMA, K. and WEIR, T. (2011). Renewable energy and climate change. In *IPCC Special Report on Renewable Energy Sources and Climate Change Mitigation* (O. Edenhofer, R. Pichs-Madruga, Y. Sokona, K. Seyboth, P. Matschoss, S. Kadner, T. Zwickel, P. Eickemeier, G. Hansen and S. Schlömer, eds.) 161–208. Cambridge Univ. Press, Cambridge.

OBAMA, B. (2017). The irreversible momentum of clean energy. *Science* **355** 126–129.

POPPICK, A. and STEIN, M. L. (2014). Using covariates to model dependence in nonstationary, high-frequency meteorological processes. *Environmetrics* **25** 293–305.

PORCU, E., BEVILACQUA, M. and GENTON, M. G. (2016). Spatio-temporal covariance and cross-covariance functions of the great circle distance on a sphere. *J. Amer. Statist. Assoc.* **111** 888–898. MR3538713

PRIESTLEY, M. B. (1965). Evolutionary spectra and non-stationary processes. *J. Roy. Statist. Soc. Ser. B* 204–237.

SOMAN, S. S., ZAREIPOUR, H., MALIK, O. and MANDAL, P. (2010). A review of wind power and wind speed forecasting methods with different time horizons. In *North American Power Symposium* (*NAPS*), 2010 1–8. IEEE.

STEIN, M. L. (2007). Spatial variation of total column ozone on a global scale. *Ann. Appl. Stat.* **1** 191–210. MR2393847

TAGLE, F., CASTRUCCIO, S., CRIPPA, P. and GENTON, M. G. (2017). Assessing potential wind energy resources in Saudi Arabia with a skew-t distribution. ArXiv Preprint arXiv:1703.04312.

TAYLOR, K. E., STOUFFER, R. J. and MEEHL, G. A. (2012). An overview of CMIP5 and the experiment design. *Bull. Am. Meteorol. Soc.* **93** 485–498.

TUKEY, J. W. (1967). An introduction to the calculations of numerical spectrum analysis. In *Advanced Seminar on Spectral Analysis of Time Series* (B. Harris, ed.) 25–46. Wiley, New York.

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. *Clim. Change* **109** 5–31.

VAUGHAN, R. A. and CRACKNELL, A. P. (2013). *Remote Sensing and Global Climate Change* **24**. Springer Science & Business Media.

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* (O. Edenhofer, R. Pichs-Madruga, Y. Sokona, K. Seyboth, P. Matschoss, S. Kadner, T. Zwickel, P. Eickemeier, G. Hansen and S. Schlömer, eds.) 535–608. Cambridge Univ. Press, Cambridge.

XU, G. and GENTON, M. G. (2017). Tukey g-and-h random fields. *J. Amer. Statist. Assoc.* **112** 1236–1249.

ZHU, X. and GENTON, M. G. (2012). Short-term wind speed forecasting for power system operations. *Int. Stat. Rev.* **80** 2–23. MR2990340

J. JEONG
M. G. GENTON
CEMSE DIVISION
KING ABDULLAH UNIVERSITY OF SCIENCE
AND TECHNOLOGY
THUWAL, 23955-6900
SAUDI ARABIA
E-MAIL: jaehong.jeong@kaust.edu.sa
marc.genton@kaust.edu.sa

S. CASTRUCCIO
DEPARTMENT OF APPLIED AND COMPUTATIONAL
MATHEMATICS AND STATISTICS
UNIVERSITY OF NOTRE DAME
NOTRE DAME, INDIANA 46556
USA
E-MAIL: scastruc@nd.edu

P. CRIPPA
DEPARTMENT OF CIVIL & ENVIRONMENTAL
ENGINEERING & EARTH SCIENCE
UNIVERSITY OF NOTRE DAME
NOTRE DAME, INDIANA 46556
USA
E-MAIL: pcrippa@nd.edu