



A Multivariate Global Spatiotemporal Stochastic Generator for Climate Ensembles

Matthew EDWARDS , Stefano CASTRUCCIO, and Dorit HAMMERLING

In order to understand and quantify the uncertainties in projections and physics of a climate model, a collection of climate simulations (an ensemble) is typically used. Given the high-dimensionality of the input space of a climate model, as well as the complex, nonlinear relationships between the climate variables, a large ensemble is often required to accurately assess these uncertainties. If only a small number of climate variables are of interest at a specified spatial and temporal scale, the computational and storage expenses can be substantially reduced by training a statistical model on a small ensemble. The statistical model then acts as a stochastic generator (SG) able to simulate a large ensemble, given a small training ensemble. Previous work on SGs has focused on modeling and simulating individual climate variables (e.g., surface temperature, wind speed) independently. Here, we introduce a SG that jointly simulates three key climate variables. The model is based on a multistage spectral approach that allows for inference of more than 80 million data points for a nonstationary global model, by conducting inference in stages and leveraging large-scale parallelization across many processors. We demonstrate the feasibility of jointly simulating climate variables by training the SG on five ensemble members from a large ensemble project and assess the SG simulations by comparing them to the ensemble members not used in training.

Supplementary materials accompanying this paper appear online.

Key Words: Nonstationary; Massive data; Stepwise estimation; Parallel computation.

1. INTRODUCTION

Climate models are fundamental in providing physically constrained projections of future climate in the geoscience community (e.g., Pachauri et al. (2014)). Despite their ever-increasing complexity, climate models are nevertheless imperfect representations of the climate system, both because of their approximate description of the physics and the intrinsic uncertainty about future emission scenarios. A large collection, or ensemble, of simulations

Matthew Edwards (✉), School of Mathematics, Statistics and Physics, Newcastle University, Newcastle upon Tyne NE1 7RU, UK (E-mail: m.edwards3@ncl.ac.uk). Stefano Castruccio, Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN 46556, USA (E-mail: scastruc@nd.edu). Dorit Hammerling, Department of Analytics and Integrative Machine Learning, National Center for Atmospheric Research, Boulder, CO 80307, USA (E-mail: dorith@ucar.edu).

© 2019 The Author(s)

Journal of Agricultural, Biological, and Environmental Statistics, Volume 24, Number 3, Pages 464–483
<https://doi.org/10.1007/s13253-019-00352-8>

is typically used to understand and quantify these uncertainties, although the computational and storage costs associated with such an endeavor are substantial. For instance, the sixth Coupled Model Inter-comparison Project (CMIP6) (Meehl et al. 2014) is expected to require approximately 1 billion core-hours of computation and 12 PB of storage (Paul et al. 2015). At the National Center for Atmospheric Research (NCAR) Supercomputing Center, this corresponds to approximately 290 compute days and one-fifth of the file storage system.

While the size of data produced by modern ensembles is significant, often only a small number of climate variables are of interest to a climate scientist. For instance, between June 2014 and March 2018, approximately 64% of the 1,168 climate variables from the Community Earth System Model Large Ensemble (CESM-LE, (Kay et al. 2015)) had never been downloaded and approximately 14% of the climate variables contribute to over 90% of the 178,121 downloads (Strand and Baker 2018). Since only a small number of climate variables are used by large parts of the research community, it is reasonable to focus efforts on developing methods to effectively reduce the computational and storage expenses for these high-interest variables while preserving their scientific integrity.

Compressing climate model output is a topic which has received increasing interest from the climate community (Baker et al. 2014, 2016). Statistical methodologies have concurrently emerged in recent years to provide accurate stochastic approximations of ensemble members, either conditionally (Guinness and Hammerling 2018) or unconditionally (Jeong et al. 2018). Jeong et al. (2018, 2019) proposed a statistical model, i.e., a stochastic generator (SG), which provides an approximation of the climate simulations, trained on a small ensemble to simulate more members. The parameters estimated from the SG are then regarded as all the information necessary to simulate the model, thus significantly reducing the information necessary to generate surrogate simulations.

SGs are models loosely inspired from stochastic weather generators (see Porcu et al. (2019) for a complete discussion on their relationship). SGs have been so far mostly focused on global climate ensembles, by proposing global spatiotemporal (GST) statistical models to capture spatiotemporal dependencies across the globe. A wealth of literature has been developed in recent years on modeling global data; see, e.g., Hitzenko and Stein (2012) for a basis decomposition approach, Jun and Stein (2007, 2008) for a constructive approach for univariate fields and Jun (2011) for the only (at present) approach for multivariate fields. In the context of global climate ensembles, Castruccio and Stein (2013) proposed a model for annually averaged surface temperature over a one-degree latitude–longitude grid that was able to capture latitudinal nonstationarities and computationally scale to higher grid resolutions. Castruccio and Genton (2014) then proposed a nonparametric generalization to capture longitudinal nonstationarities and later Castruccio and Guinness (2017) proposed a model to account for land/ocean dependence. Castruccio and Genton (2016) generalized these models for vertical temperatures, and Castruccio (2016) to multiple climate models. Recently, Castruccio and Genton (2018) provided a general theory for the aforementioned class of models, and Castruccio et al. (2019) discussed implications on visualization. All these works rely on a parametric approach, since nonparametric alternatives such as principal component analysis would be burdensome in terms of storage, and the aim is to achieve the highest possible compression rate.

All the aforementioned GST models, with the exception of Jun (2011), have been used as SGs for a single variable such as temperature or wind speed. While diagnostics have demonstrated the suitability of GST models in representing the spatiotemporal dependencies for a single variable, they have not been designed to capture multivariate dependencies. A natural question is therefore if and to what extent a multivariate SG can be formulated to jointly simulate variables of interest. A multivariate SG is of high scientific interest since the dependencies between the climate variables are often the primary focus. Indeed, the geoscience community is primarily concerned in integrated assessments by describing interactions among key climate variables such as temperature and precipitation, as well as their variability among ensemble members (e.g., Pachauri et al. (2014)).

For this purpose, we develop a multivariate extension of the GST model described in Castruccio and Stein (2013). This generalization relies on a novel multivariate spectral model for data on the sphere and in time, whose implied marginal models correspond to the GST model previously available in the literature. The model has a marginally parameterized (MP) structure (Edwards et al. 2018), which allows for multistage estimation for millions of data points. Inference is performed in four steps: temporal, longitudinal, latitudinal and multivariate. Our proposed model naturally generalizes the existing literature by adding dependence across variables to the model formulation.

We develop univariate and multivariate SGs for three commonly downloaded annually averaged variables from the CESM-LE (Kay et al. 2015): total precipitable water, surface temperature and 10 m wind speed. Total precipitable water was selected over precipitation since it avoids the need for zero-inflated models to account for the high proportion of zero values and a Gaussian assumption is more appropriate. The CESM ensemble is divided into a training ensemble and a test ensemble; the former is used to train both SGs. For validation, we compare simulations from the univariate and multivariate SGs to the test ensemble.

The remainder of this paper is organized as follows: In Sect. 2, we introduce the CESM-LE and the three climate variables. The notation for the test ensemble and the multivariate GST model for the SG are provided in Sect. 3, and the definition of the MP model and the multistage estimation method are given in Sect. 4. In Sect. 5, we present the results of simulating the three variables independently and jointly and conclude in Sect. 6.

2. CESM LARGE ENSEMBLE

2.1. DATASET EXPLORATORY ANALYSIS

The CESM-LE (Kay et al. 2015) is a publicly available (<https://www.earthsystemgrid.org/>) collection of climate system simulations intended to understand uncertainties related to climate variability and climate change. The CESM-LE consists of 33 members simulated from 1920 to 2100 on an approximately one-degree latitude–longitude grid using the Community Atmosphere Model version 5 (CAM5) (Hurrell et al. 2013). The ensemble members are produced through round-off-level perturbations of the initial atmospheric temperature field. Due to the chaotic nature of the climate system—a nonlinear dynamical system—and its sensitivity to initial conditions, the simulations are approximately independent and identically distributed (i.i.d.) (Collins and Allen 2002; Collins 2002; Branstator and Teng 2010).

The historical greenhouse gas concentration pathway was used from 1920 to 2005 and the Representative Concentration Pathway (RCP) 8.5 (Moss et al. 2008) from 2006 to 2100. Further details of the model setup can be found in Kay et al. (2015).

Based on consultations with climate scientists, we select three commonly downloaded climate variables with important multivariate dependencies for the SG: total precipitable water (TMQ), surface temperature (TS) and 10 m wind speed (U10). TMQ is the mass of water in a column of the atmosphere per unit area (kg/m^2), TS is the radiative surface temperature (K) and U10 is the wind speed 10 m above the surface (m/s). These climate variables are annually averaged to approximate, through the central limit theorem, the Gaussian assumption specified in Sect. 3. Higher temporal resolutions would require non-Gaussian models for our data. To reduce the substantial computational and memory expenses required to train the SG, we only consider the 2006 to 2100 time period. This setup corresponds to 33 members, $T = 95$ years, $L = 288$ longitudes, $M = 192$ latitudes and $V = 3$ variables; a total of approximately half a billion data points. In Jeong et al. (2018), it was concluded, with a lack of fit metric, that five randomly selected members were sufficient for training a model with similar temporal, longitudinal and latitudinal specifications. Since the number of training ensemble members is the number of ensemble members simulated in practice, no more members than necessary are selected for the training ensemble. Consequently, the 33 ensemble members are divided into a training ensemble containing $R = 5$ randomly selected members (as suggested in Jeong et al. (2018)) and a test ensemble contains the 28 remaining members. The training ensemble corresponds to approximately 80 million data points. The training ensemble is used to train the SG, and the test ensemble is used to diagnose the quality of the SG simulations.

To obtain an understanding of the differences between the selected climate variables, we fit a simple linear regression model, with year as the regressor, at each spatial location for each variable (Fig. 1). The intercepts (residual standard deviations) are larger (larger) around the tropics for TMQ, larger (smaller) toward low latitudes for TS and larger (larger) over the ocean for U10. Without going into details, it is interesting to note that the primary sources for these variations are quite different: For TMQ, it is the Hadley cell; for TS, the solar zenith angle; and for U10, the surface type (Washington and Parkinson 2005, Chapter 2).

3. STATISTICAL MODEL

3.1. NOTATION

Let Y denote the $R \times T \times L \times M \times V$ training ensemble array where each element $Y[r, t, l, m, v]$ corresponds to a member $r = 1, \dots, R$, a year $t = 1, \dots, T$, a longitude $l = 0, \dots, L - 1$, a latitude $m = 1, \dots, M$ and a variable $v = 1, \dots, V$. Zero-based indexing for longitude is adopted for the spectral methods introduced in Sects. 3.6 and 3.7. Indices from the notation $Y[r, t, l, m, v]$ are omitted to denote subarrays of Y . For example, $Y[r, t, v]$ denotes the $L \times M$ matrix indexed over longitude (rows) and latitude (columns) for member r , year t and variable v . Therefore, $Y_r := Y[r, \cdot]$ is the $T \times L \times M \times V$ subarray

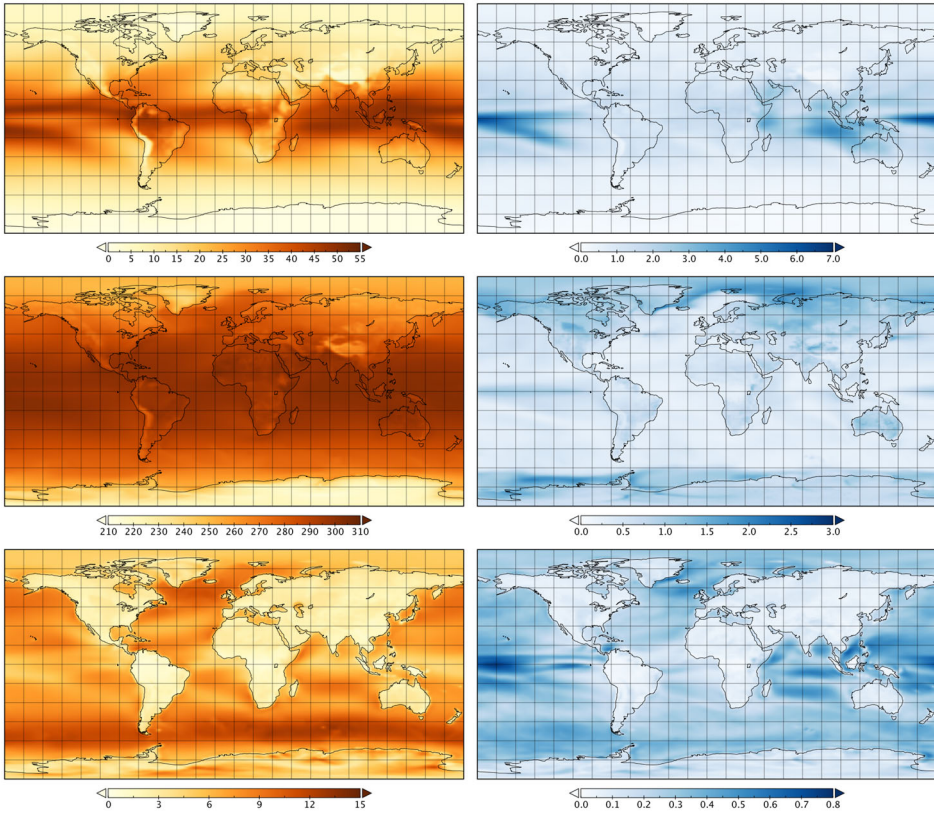


Figure 1. These plots highlight the structural differences between the TMQ (kg/m^2), TS (K) and U10 (m/s) climate variables. Intercepts (left panels) and residual standard deviations (right panels) from simple linear regression models, with year as the regressor, fit at each spatial location for TMQ (upper panels), TS (middle panels) and U10 (lower panels) from the CESM-LE..

for member r . For modeling, define the training ensemble vector as $\mathbf{y} := \text{vec}(\mathbf{Y})$ where $\text{vec}(\cdot)$ is the vec-operator

$$\text{vec}(\mathbf{Y}) := \sum_{v \in \mathbb{Z}_V} \cdots \sum_{r \in \mathbb{Z}_R} (\mathbf{e}_{v,V} \otimes \cdots \otimes \mathbf{e}_{r,R}) \cdot \mathbf{Y}[r, t, l, m, v],$$

where $\mathbf{e}_{k,K}$ is a unit vector of order K (see Hardy and Steeb 2010). Additionally, let the subarray notation applied to the vector \mathbf{y} be defined as the subarray notation applied to \mathbf{Y} followed by the vec-operator. For example, the time series vector corresponding to member r , longitude l , latitude m and variable v is denoted $\mathbf{y}_r[l, m, v]$ where $\mathbf{y}_r := \text{vec}(\mathbf{Y}_r)$. In general, uppercase letters denote arrays, bold lowercase letters denote vectors and bold uppercase letters denote random vectors.

3.2. ENSEMBLE MODEL

It is assumed that the ensemble members \mathbf{Y}_r for $r \in \mathbb{Z}_R$ are i.i.d. as

$$\mathbf{Y}_r \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\Sigma}$ is the covariance matrix. Since the variables are annually averaged, the Gaussian assumption is approximated as a result of the central limit theorem. Due to the small training ensemble ($R = 5$), in contrast to Castruccio and Stein (2013), Castruccio and Genton (2016) and Castruccio and Guinness (2017), we do not estimate $\boldsymbol{\mu}$ nonparametrically using a restricted likelihood (Patterson and Thompson 1971). Instead, $\boldsymbol{\mu}$ is incorporated parametrically into the temporal model specified in Sect. 3.3.

3.3. TEMPORAL MODEL

The training ensemble members \mathbf{Y}_r for $r \in \mathbb{Z}_R$ are modeled with a vector auto-regressive moving average model (Lütkepohl 2005, Chapter 11) with diagonal auto-regressive and moving average matrices (DVARMA). The auto-regressive (AR) and moving average (MA) orders are p and q , respectively. The DVARMA model is specified as

$$\mathbf{Y}_r[t] = \sum_{i=0}^d \boldsymbol{\beta}_i x_{i,t} + \sum_{i=1}^p \Phi_i \mathbf{Y}_r[t - i] + \sum_{i=0}^q \Pi_i \mathbf{S} \mathbf{H}_r[t - i], \tag{1}$$

where $\boldsymbol{\beta}_i$ are vectors of mean parameters, \mathbf{x}_i are the normalized orthogonal columns of the Q matrix from the QR decomposition of the $T \times d$ Vandermonde matrix $V_{t,j} := t^{j-1}$ (Golub and Van Loan 2012, Sect. 4.6), $\Phi_i := \text{diag}(\phi_{i,l,m,v})$ are diagonal matrices of AR parameters, $\Pi_i := \text{diag}(\pi_{i,l,m,v})$ are diagonal matrices of MA parameters where $\pi_{0,l,m,v} = 0$, $\mathbf{S} := \text{diag}(s_{l,m,v})$ are diagonal matrices of standard deviation parameters and $\mathbf{H}_r[t] \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$ are i.i.d. zero-mean and unit-variance multivariate spatial innovations with correlation matrix \mathbf{R} . Autocorrelation function plots suggest that the multivariate spatial residuals are uncorrelated, supporting the model specification; see Fig. S1 in supplemental material. This DVARMA model can represent variation in mean, trend, variance and temporal correlation over space and climate variables. The DVARMA model cannot, however, represent variation in phase over space and climate variables.

3.4. SPECTRAL METHODS

Sections 3.6 and 3.7 require an understanding of spectral methods. Consider a zero-mean and unit-variance stationary discrete Gaussian processes $Z[s]$ for $s = 1, \dots, S$. In the spatial (or temporal) domain, the dependence between these random variables is described with a correlation function $C(h) := \text{corr}(Z(s+h), Z(s))$. In the spectral domain, the stationary discrete Gaussian process is considered as a sum of complex exponential functions

$$Z[s] = \sum_{c=0}^{S-1} Z[c] \exp(2\pi i s c / S),$$

where $Z[c]$ are independent zero-mean complex-valued random variables. The value c corresponds to a wave number since the complex exponential function corresponding to c is a sum of sinusoidal functions with frequency c/K . Since the complex-valued random variables are independent, they are described not by correlation but variance $f(c) := \text{var}(Z[c])$. The

variances are called the spectral mass function (SMF). The SMF describes how a stationary Gaussian process is decomposed into different frequency sinusoidal functions. The extension to multiple zero-mean and unit-variance stationary discrete Gaussian processes is similar.

3.5. INNOVATION MODEL

Let $\mathbf{H}_{r,t} := \mathbf{H}_r[t]$ denote the i.i.d. zero-mean and unit-variance multivariate spatial innovations for member r and year t . It is assumed that

$$\text{corr}(\mathbf{H}_{r,t}[l+h, m_1, v_1], \mathbf{H}_{r,t}[l, m_2, v_2]) = C(h, m_1, m_2, v_1, v_2),$$

where h is the longitudinal lag and $C(\cdot)$ is a positive definite multivariate spatial cross-correlation function. This assumption is a multivariate extension of axial symmetry (Jones 1963) and is the multivariate analogue to the altitudinal extension introduced in Castruccio and Genton (2016). Given the multivariate axial symmetry assumption, all the longitudinal bands are jointly stationary. Let $\mathbf{R}_{m_1, m_2, v_1, v_2}$ denote the cross-correlation matrix between the longitudinal bands $\mathbf{H}_{r,t}[m_1, v_1]$ and $\mathbf{H}_{r,t}[m_2, v_2]$. Since the longitudinal bands are circular, the cross-correlation matrices are circulant (Davis 2012). Hence, $\mathbf{R}_{m_1, m_2, v_1, v_2}$ can be decomposed as

$$\mathbf{R}_{m_1, m_2, v_1, v_2} = \mathbf{W}^{-1} \text{diag}(\mathbf{W} \mathbf{r}_{m_1, m_2, v_1, v_2}) \mathbf{W}, \quad (2)$$

where \mathbf{W} is the $L \times L$ discrete Fourier transform matrix, with element $\exp(2\pi i j k / L)$ in row j and column k where $i := \sqrt{-1}$, and $\mathbf{r}_{m_1, m_2, v_1, v_2}$ is the first column of $\mathbf{R}_{m_1, m_2, v_1, v_2}$ (Davis 2012, Sect. 3.2). The vector $\mathbf{W} \mathbf{r}_{m_1, m_2, v_1, v_2}$ defines the cross-spectral mass function f_{m_1, m_2, v_1, v_2} , which can be decomposed as

$$f_{m_1, m_2, v_1, v_2} = f_{m_1, v_1}^{1/2} \cdot f_{m_2, v_2}^{1/2} \cdot \rho_{m_1, m_2, v_1, v_2} \cdot \exp(i \phi_{m_1, m_2, v_1, v_2}), \quad (3)$$

where $f_{m, v}$ is the SMF, $\rho_{m_1, m_2, v_1, v_2}$ is the coherence mass function (CMF) and $\phi_{m_1, m_2, v_1, v_2}$ is the phase mass function (PMF). These functions, which will be specified in Sects. 3.6 and 3.7, fully characterize a model for the multivariate spatial innovations given the multivariate axially symmetric assumption.

3.6. LONGITUDINAL MODEL

The longitudinal model controls the correlation within longitudinal bands. Since $\mathbf{R}_{m, m, v, v}$ is a correlation matrix, where $\rho_{m, m, v, v} \equiv 1$ and $\phi_{m, m, v, v} \equiv 0$, the longitudinal model only requires the specification of the SMFs. Furthermore, since the multivariate spatial innovations are unit variance

$$\sum_{c \in \mathbb{Z}_L} f_{m,v}(c) = L, \tag{4}$$

see ‘‘Appendix A’’ for a proof, the SMFs do not require scale parameters and only require specification up to a constant of proportionality. Since scale parameters are included in Castruccio and Stein (2013), Castruccio and Genton (2016) and Castruccio and Guinness (2017), relationship (4) resulted in a reduction in parameters. The Modified Matérn SMF introduced in Castruccio and Stein (2013) is a Matérn SMF modified for circular processes. It assumes that there is a smooth spectrum transition at high frequencies. This assumption is appropriate for TS; however, it is not appropriate for TMQ and U10. We propose a γ -Modified Matérn SMF, with a parameter that controls the spectrum transition at high frequencies. The γ -Modified Matérn SMF is defined as

$$f_{m,v}(c) \propto \frac{1}{(\alpha_{m,v}^2 + \gamma_{m,v} A^2(c) + (1 - \gamma_{m,v}) B^2(c))^{\kappa_{m,v} + 1/2}}, \tag{5}$$

where

$$A(c) := 2 \sin(\pi c/L), \quad B(c) := 2(1 - |2c/L - 1|).$$

$\alpha_{m,v}$ is the inverse range parameter, $\gamma_{m,v}$ is the transition parameter and $\kappa_{m,v}$ controls the increased decay rate in spectral mass for larger wave numbers. These parameters are allowed to vary over variables and latitudes. When $\gamma_{m,v} = 1$, the Modified Matérn SMF is recovered. Note that this is a Whittle model (Whittle 1954).

3.7. LATITUDINAL AND MULTIVARIATE MODEL

The latitudinal and multivariate model controls the correlation between longitudinal bands. Since $\mathbf{R}_{m_1, m_2, v_1, v_2}$ is a cross-correlation matrix, which depends on $\rho_{m_1, m_2, v_1, v_2}$ and $\phi_{m_1, m_2, v_1, v_2}$, the latitudinal and multivariate model requires the specification of the CMFs and the PMFs. The CMFs and PMFs are specified implicitly through a dynamical model. The dynamical model allows for efficient simulation; see Sect. 5 of Castruccio and Guinness (2017) for details. Let $\tilde{\mathbf{H}}_{r,t}[m, v]$ denote the discrete Fourier transform of $\mathbf{H}_{r,t}[m, v]$. The Fourier coefficients are modeled with a diagonal vector AR (DVAR) model of order one (Jeong et al. 2018). The DVAR model is specified as

$$\tilde{\mathbf{H}}_{r,t}[c, m] = \Psi_{c,m} \tilde{\mathbf{H}}_{r,t}[c, m - 1] + \epsilon_{r,t}[c, m],$$

where $\tilde{\mathbf{H}}_{r,t}[c, m]$ are vectors of Fourier coefficients, $\Psi_{c,m} := \text{diag}(\psi_{c,m,1}, \dots, \psi_{c,m,v})$ are diagonal matrices of AR parameters and $\epsilon_{r,t}[c, m] \sim \mathcal{N}(\mathbf{0}, \Xi_{c,m})$ are zero-mean and unit-variance multivariate innovations with correlation matrices $\Xi_{c,m}$ independent in r and t . Under the constraint that

$$\Xi_{c,m}[v_1, v_2] := \Xi_c[v_1, v_2] (1 - \psi_{c,m,v_1} \psi_{c,m,v_2}),$$

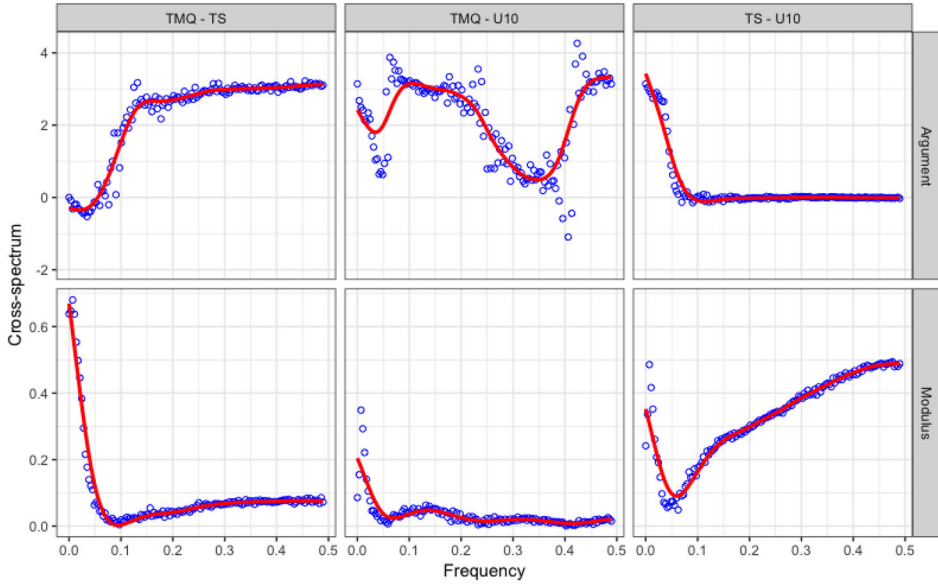


Figure 2. Moduli and arguments of cross-periodograms (blue) and the natural cubic splines (red) between the TMQ and TS climate variables, the TMQ and U10 climate variables and the TS and U10 climate variables over frequency (Color figure online)..

where $\Xi_c := \Xi_{c,0}$ (this constraint implies that the multivariate spatial innovations have unit variance), the cross-spectral mass functions are

$$f_{m_1, m_2, v_1, v_2}(c) = \Xi_c[v_1, v_2] \left(\prod_{j=m_1+1}^{m_2} \psi_{c,j,v_2} \right), \tag{6}$$

for $m_1 < m_2$, m_1 and m_2 are exchanged for $m_2 < m_1$, see ‘‘Appendix B’’ for a proof. For $m_1 = m_2$ the purely multivariate dependence is controlled by Ξ_c only and for $v_1 = v_2$ the purely latitudinal dependence is controlled by the AR parameters only. Furthermore, the purely latitudinal dependence is the same as that in Castruccio and Guinness (2017). Following Castruccio and Guinness (2017), the AR parameters are parameterized as

$$\psi_{c,m,v} = \delta_{v,m} \left(1 + 4 \sin^2 \frac{\pi c}{L} \right)^{-\tau_{m,v}}, \tag{7}$$

where $\delta_{v,m}$ controls the rate of decay in coherence, over all wave numbers, as the distance between latitudes increases and $\tau_{v,m}$ controls the increased decay rate in coherence for larger wave numbers. These parameters are allowed to vary over variables and latitudes. The multivariate model is fully specified by Ξ_c . Since the cross-periodograms between the climate variables are not easily captured with a parametric model (see Fig. 2), for each climate variable the moduli and arguments of $\Xi_c[v_1, v_2]$ are modeled with a natural cubic spline (Friedman et al. 2001, Sect. 5.2.1) over the wave numbers. Note that $\Xi_c[v_1, v_2] = \Xi_c[v_2, v_1] = 0$ implies that the climate variables v_1 and v_2 are independent.

One might consider increasing the flexibility of this model by allowing the innovations to have different correlations over latitudes. However, to constrain the multivariate spatial innovations to have unit variance with this additional flexibility is a non-trivial problem to be addressed in future work.

4. ESTIMATION

In this section, we prove that the parameters of the model introduced in Sect. 3 can be estimated with a sequence of marginal likelihood functions. This property of estimation defines a marginally parameterized (MP) model (Edwards et al. 2018) and allows the parameters to be estimated efficiently in multiple stages.

4.1. MARGINALLY PARAMETERIZED MODEL

Heuristically, a model for \mathbf{y} is MP if there exists a finite sequence of data subsets such that the corresponding sequence of marginal likelihood functions can be used to estimate all of the model parameters.

Definition 1. (Marginally parameterized model) A model for \mathbf{y} is marginally parameterized if there exists a finite sequence of $K > 1$ data subsets (\mathbf{y}_k) such that the marginal model of \mathbf{y}_k depends on a parameter subset with a partition $\boldsymbol{\theta}_k, \boldsymbol{\eta}_k$ where $\boldsymbol{\theta}_k \neq \emptyset$ and $\boldsymbol{\eta}_k \subseteq \boldsymbol{\theta}_1 \cup \dots \cup \boldsymbol{\theta}_{k-1}$ ($\boldsymbol{\eta}_1 = \emptyset$) for $k = 1, \dots, K$ and $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K$ is a partition of $\boldsymbol{\theta}$.

Consider the finite sequence of $V + 1$ data subsets $\mathbf{y}_v := \mathbf{y}[v]$ for $v = 1, \dots, V$ and $\mathbf{y}_{V+1} := \mathbf{y}$. Since the multivariate GST model is Gaussian and $f_{m_1, m_2, v, v}(c)$ is the cross-spectral mass function for the univariate GST model (see Sect. 3.7), the marginal model of \mathbf{y}_v is the univariate GST model. As a consequence, the univariate GST likelihood function can be used to estimate the univariate (temporal, longitudinal and latitudinal) parameters for each variable $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_V$ and the multivariate GST likelihood function can be used to estimate the multivariate parameters $\boldsymbol{\theta}_{V+1}$ conditional on the estimates for the univariate parameters $\boldsymbol{\eta}_{V+1} = \boldsymbol{\theta}_1 \cup \dots \cup \boldsymbol{\theta}_V$. Hence, the multivariate GST model is MP. Theorem 3.1 in Edwards et al. (2018) provides the conditions under which these estimates are consistent.

4.2. MULTISTAGE ESTIMATION

The parameters of the multivariate GST model introduced in Sect. 3 can be estimated in four stages. In these four stages, the temporal, longitudinal, latitudinal and multivariate parameters are estimated, respectively, with each stage conditioning on parameters estimated in previous stages.

- **Stage one** The dataset is partitioned into $L \cdot M \cdot V$ data subsets $\mathbf{y}[l, m, v]$. The marginal model for data subsets is a product of R ARMA models. The ARMA model primary (temporal) parameters for $\mathbf{y}[l, m, v]$ are $\beta_{0,l,m,v}, \dots, \phi_{1,l,m,v}, \dots, \pi_{1,l,m,v}, \dots, s_{l,m,v}$ (1). The computational and memory cost of evaluating each ARMA model is $\mathcal{O}(T)$

and $\mathcal{O}(T)$, respectively. Selection of d , p and q is performed with AIC. Conditioning on the temporal parameters results in multivariate spatial residuals \mathbf{h} .

- **Stage two** The multivariate spatial residuals are partitioned into $M \cdot V$ data subsets $\mathbf{h}[m, v]$. The marginal model for $\mathbf{h}[m, v]$ is a product of $R \cdot T$ Whittle models. The Whittle model primary (longitudinal) parameters for $\mathbf{h}[m, v]$ are $\alpha_{m,v}, \gamma_{m,v}, \kappa_{m,v}$ (5). The computational and memory cost of evaluating each Whittle model is $\mathcal{O}(L \ln(L))$ and $\mathcal{O}(L)$, respectively. Selection of the modified or γ -modified Matérn is performed with AIC.
- **Stage three** The multivariate spatial residuals are partitioned into V data subsets $\mathbf{h}[v]$. The marginal model for $\mathbf{h}[v]$, conditional on the longitudinal parameters, is a product of $R \cdot T \cdot L$ Gaussian models. The Gaussian model primary (latitudinal) parameters for $\mathbf{h}[v]$ are $\delta_{1,v}, \dots, \delta_{M,v}, \tau_{1,v}, \dots, \tau_{M,v}$ (6). The computational and memory cost of evaluating each Gaussian model is $\mathcal{O}(M^3)$ and $\mathcal{O}(M^2)$, respectively. Selection of the stationary or nonstationary model is performed with AIC.
- **Stage four** The (marginal) model for \mathbf{h} , conditional on the longitudinal and latitudinal parameters, is a product of $R \cdot T \cdot L \cdot M$ Gaussian models. The Gaussian model primary parameters for \mathbf{h} are those of the natural cubic spline. The computational and memory cost of each Gaussian model is $\mathcal{O}(V^3)$ and $\mathcal{O}(V^2)$, respectively. Selection of the degrees of freedom is performed with AIC.

The collection of data subsets for each stage can be used to estimate the corresponding parameters in parallel. Hence, the parameters can be estimated very efficiently with parallel computation.

5. RESULTS

The applicability of a SG depends on how accurately its simulations can represent climate variables from an ensemble. Hence, in this section, we compare the test ensemble, consisting of three variables, to simulations from our univariate and multivariate GST models. The only difference between the univariate and multivariate GST models is that we set $\Xi[v_1, v_2] \equiv 0$ for $v_1 \neq v_2$ for the univariate SG. This condition implies that the three climate variables are simulated independently rather than jointly. Diagnostics are provided in two stages. First, we provide univariate diagnostics to assess the similarity of the SG simulations to climate model simulations using the test ensemble (members not in the training ensemble). We then provide multivariate diagnostics to assess the difference of simulating climate variables jointly rather than independently. Differences between the diagnostics are primarily gauged visually at this stage in the SG development (Castruccio et al. 2019).

5.1. UNIVARIATE DIAGNOSTICS

We consider area-weighted statistics (to account for the non-regular latitude–longitude grid with more grid points per unit area toward the poles) to compare the ensembles from the

Table 1. Mean (over years, longitudes and latitudes) and standard deviation (in parenthesis) of the weighted minimum, weighted first quartile, weighted median, weighted mean, weighted third quartile and weighted max over the test ensemble (left columns) and the joint SG (right columns) ensembles.

Ensemble	TMQ (kg/m ²)		TS (K)		U10 (K)	
	Test	Joint SG	Test	Joint SG	Test	Joint SG
Min.	0.23 (0.01)	0.18 (0.03)	214.60 (0.25)	214.50 (0.32)	0.65 (0.00)	-0.34 (0.13)
First Qu.	15.04 (0.03)	15.03 (0.02)	282.56 (0.03)	282.56 (0.02)	3.40 (0.00)	3.40 (0.00)
Median	25.31 (0.04)	25.31 (0.04)	294.90 (0.02)	294.90 (0.02)	6.53 (0.01)	6.53 (0.01)
Mean	28.05 (0.03)	28.05 (0.05)	289.64 (0.02)	289.64 (0.02)	6.12 (0.00)	6.12 (0.00)
Third Qu.	40.66 (0.05)	40.68 (0.09)	300.46 (0.02)	300.46 (0.02)	8.00 (0.01)	8.01 (0.01)
Max.	72.66 (0.70)	75.83 (2.33)	312.16 (0.23)	312.87 (0.49)	13.73 (0.09)	13.82 (0.16)

climate model and our SG (Table 1). The medians and means from the test ensemble and the joint SG ensemble are equal to two decimal places, and the first and third quartiles are equal to one decimal place. This indicates that the joint SG accurately captured the body of the test ensemble distribution. The minimums and maximums are also well represented for TS and TMQ. In the case of U10, the maximum is well modeled, but the minimum is negative, which is not physically possible. While the difference is small, it nevertheless indicates the need for an improved model which can enforce nonnegativity. A log or Box-Cox transformation (Box and Cox 1964) can remove the negative values. A transformation performed before modeling can result in biased simulations. Hence, it is important to incorporate any transformation into the model to account for biases. This is a potential improvement we postpone for future work.

For succinctness, the remaining univariate diagnostics focus on the TS climate variable, since they are very similar for the TMQ and U10 climate variables. We consider global and longitudinal TS means to assess how accurately the multivariate SG captures variations in TS mean over years and latitudes, respectively (Fig. 3). The TS means overall have similar structures and variability. Upon careful inspection, the global TS means for the SG ensemble

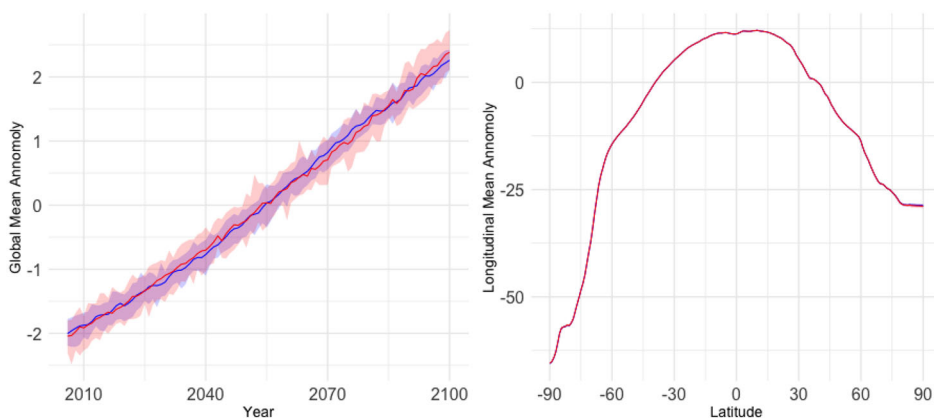


Figure 3. Global TS (K) means over time (left panel), the longitudinal TS (K) means over latitude (right panel) over the test ensemble (red) and the joint SG ensemble (blue). The maximum and minimum member global TS means, for each year, and longitudinal means, for each latitude, are included as red and blue shaded regions, respectively. Note that the longitudinal means (right panel) overlap almost entirely. (Color figure online)..

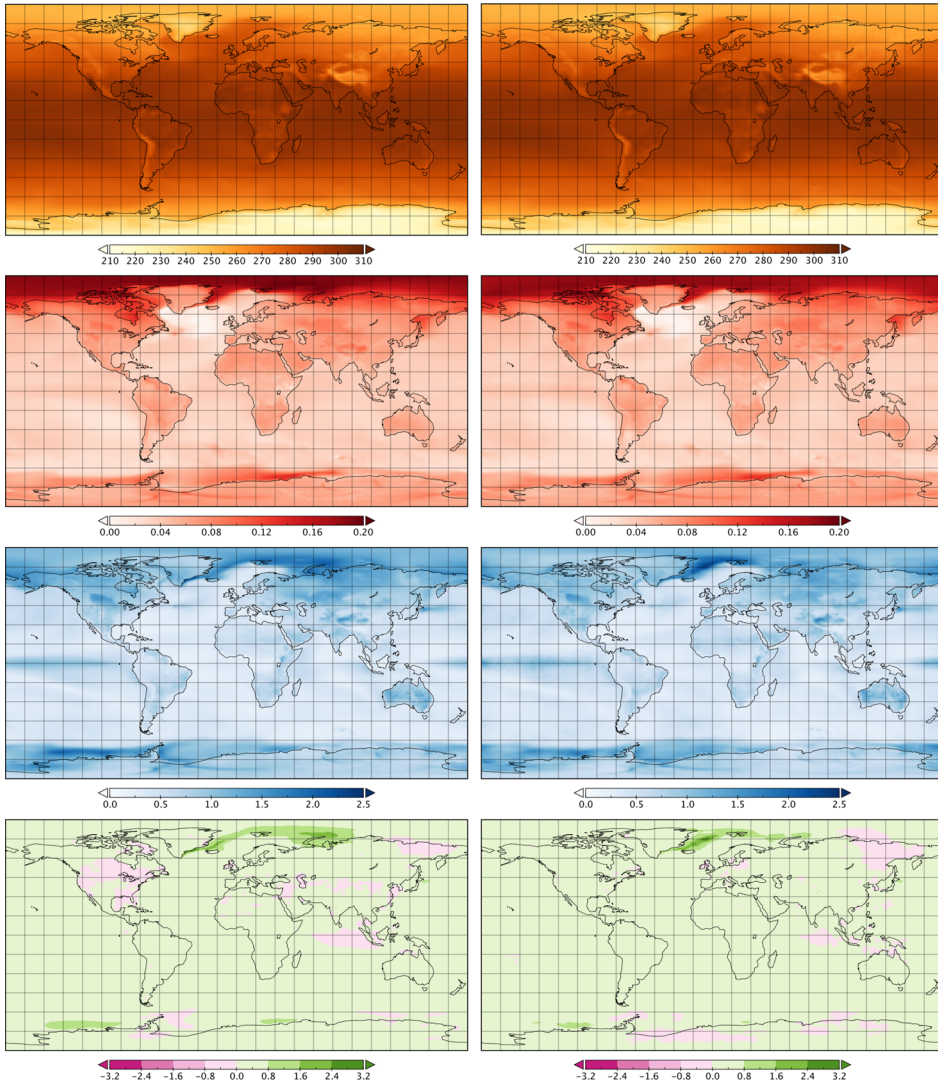


Figure 4. Intercepts (first row), the slopes (second row), the residual standard deviations (third row) and the residual lag-one auto-covariances (fourth row) from simple linear regression models trained at each spatial location for TS (K) over the test ensemble (left column) and the joint SG ensemble (right column)..

have slightly larger variability between members and their trend is also slightly larger toward 2100.

The remaining univariate and multivariate diagnostics are based on a simple linear regression model, with year as the regressor, trained at each spatial location for each variable over the test ensemble and joint SG ensemble. The intercepts, slopes, residual standard deviations and residual lag-one auto-covariances from these models are used to assess how accurately the means, trends, standard deviations and temporal auto-covariances over space are captured by the multivariate SG (Fig. 4). Figure 4 displays intercepts, slopes, residual standard deviations and residual lag-one auto-covariances corresponding to the TS climate variable. The differences between intercepts, slopes, residual standard deviations and residual lag-one

auto-covariances from the test ensemble and the joint ensemble for all the climate variables are displayed in Figs. S2, S3 and S4 in supplemental material.

The intercepts and slopes demonstrate how the multivariate GST model can capture the variation in mean and trend over space quite accurately, even though they were estimated independently over space. There are, however, some differences between the residual standard deviations. First, the standard deviations for the joint SG ensemble are slightly smaller over parts of the Arctic Ocean and the Antarctic coast than for the test ensemble. Second, in the central Pacific Ocean, the variations in standard deviations for the joint SG ensemble are less smooth than for the test ensemble. The latter suggests that the multivariate GST model could benefit from model specifications imposing smoothness in this region rather than estimating each grid cell independently over space. This region is associated with the non-periodic El-Niño-Southern Oscillation, which is notoriously difficult to model. Third, there are some differences between the residual lag-one auto-covariances. The residual lag-one auto-covariances for the joint SG ensemble over the Arctic Ocean above Russia and across the Antarctic coast below the Pacific Ocean are substantially smaller than for the test ensemble. Both the smaller residual standard deviations and residual lag-one auto-covariances occur toward the poles. This suggests the multivariate GST model is not capturing the poles as accurately as the low and middle latitudes.

5.2. MULTIVARIATE DIAGNOSTICS

To assess how accurately the univariate and multivariate SGs capture the dependencies between the three climate variables, the cross-correlation between the residuals at each

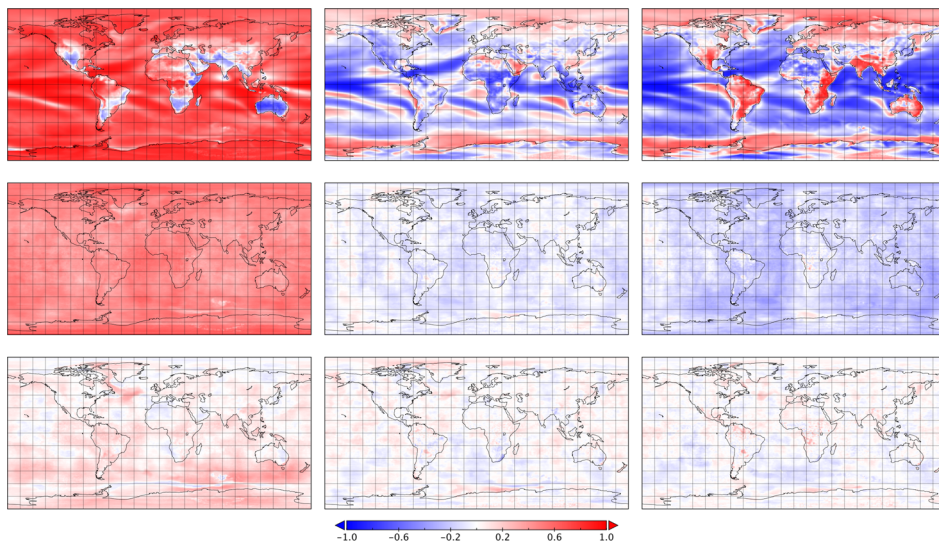


Figure 5. Cross-correlation between the TMQ and TS residuals (left panels), the TMQ and U10 residuals (middle panels) and the TS and U10 residuals (right panels) from simple linear regression model fits at each spatial location over the test ensemble (upper panels), the joint SG ensemble (middle panels) and the independent SG ensemble (lower panels)..

spatial location is displayed for each pair of variables for the test ensemble, the joint SG ensemble and the independent SG ensemble (Fig. 5). TMQ and TS (upper left panel) display strong positive and negative spatially varying cross-correlation. The lower left panel (Fig. 5) displays approximately zero cross-correlation between the TMQ and TS residuals over space. This is expected, since the TMQ and TS climate variables were simulated independently. There appears to be some structure to the slightly positive and negative cross-correlations between the TMQ and TS residuals; however, these do not correspond to the structures displayed in the upper left panel (Fig. 5). The middle left panel (Fig. 5) displays positive cross-correlation between the TMQ and U10 residuals over space. The left panels (Fig. 5) suggest that the multivariate GST model cannot capture variation in cross-correlation over space, but if there is an average positive or negative cross-correlation over space, the model will capture it. The upper middle panel (Fig. 5) suggests that the average cross-correlation between the TMQ and U10 residuals over space is approximately

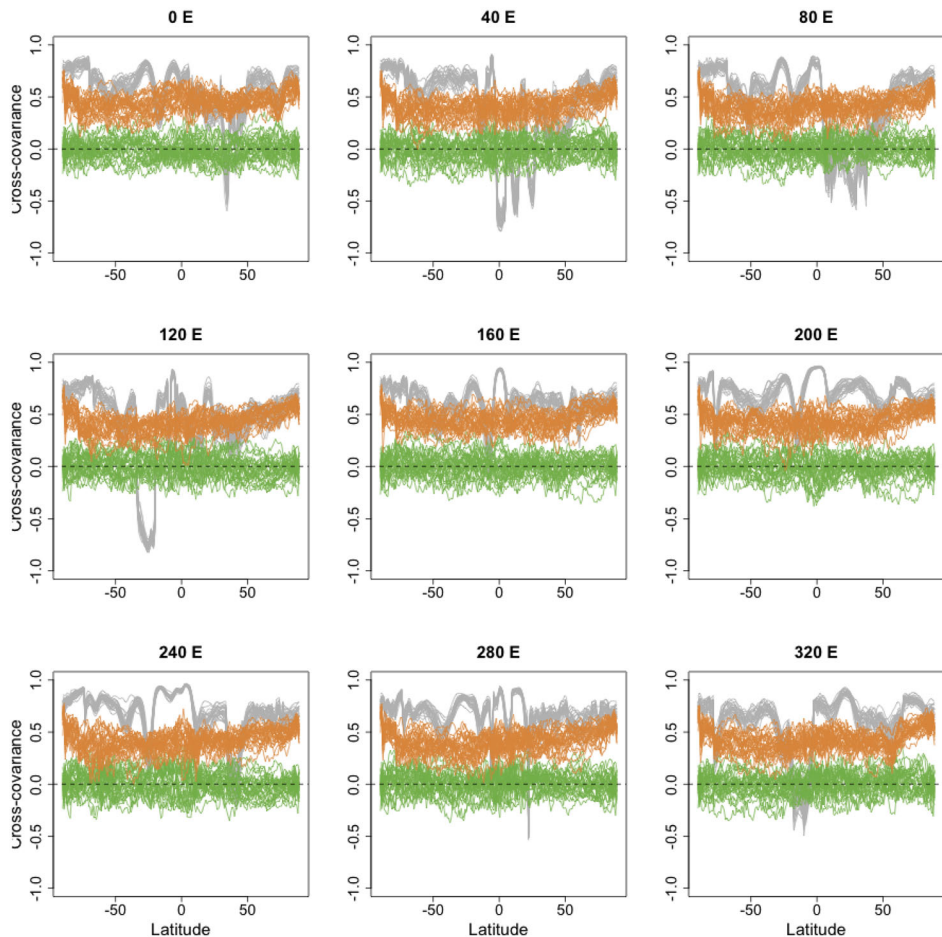


Figure 6. Cross-correlation between the TMQ and TS residuals at nine different longitudes and all latitudes for each member of the test ensemble (gray), the independent SG ensemble (green) and the joint SG ensemble (orange) (Color figure online).

zero. As suggested from the left column of panels (Fig. 5), the model can only capture the average positive or negative cross-correlation over space and not the spatial variation. However, since the average is approximately zero in the upper middle panel (Fig. 5), there is no substantial improvement to the joint SG ensemble over the independent SG ensemble. In contrast to the middle column of panels (Fig. 5), the average cross-correlation between the TS and U10 residuals is negative for the middle right panel (Fig. 5). Hence, the model has captured the average negative cross-correlation between the TS and U10 residuals over space, but not the structure displayed in the upper right panel (Fig. 5).

To assess the variation in cross-correlation over space between ensemble members, cross-correlations for each member are displayed at nine different longitudes and all latitudes (Fig. 6). These plots demonstrate how the cross-correlation of the joint SG tends toward the average cross-correlation of the test ensemble, while the cross-correlation of the univariate SG hovers around zero as expected. These figures also suggest that the variation in cross-correlation between ensemble members is smaller for the test ensemble than for the joint or independent SG ensembles.

6. CONCLUSION

For a small number of variables, a multivariate SG can be trained on a climate model ensemble consisting of only a few members to simulate a large ensemble. As a consequence, if only a small number of climate variables are of interest, then a large ensemble of these variables can be obtained with a SG with reduced computational and storage expenses. The computational expense is reduced since only a small ensemble is required to train the SG, and the storage expense is reduced since only the SG requires storage; both are important considerations for climate modeling centers.

However, the applicability of a SG depends on how accurately its simulations can represent ensemble features of interest. For example, the univariate diagnostics suggest that extreme value analysis and polar region analysis would not be appropriate for the proposed SG since extremes and polar regions are less accurately captured, while analyses that require the accurate representation of means, trends, variances and temporal autocorrelations over space would be suitable. A SG for extreme value analysis would require a non-Gaussian model. In general, the study of non-Gaussian MP models would be very important in this regard. However, since the Gaussian MP models studied so far extensively exploit closure under marginalization, the class of models that are amenable to this approach could be limited by this property. The multivariate diagnostics demonstrate that the multivariate SG can capture cross-correlations, but it is limited to an average over space. While this is a substantial improvement over univariate SGs, it still limits the application of multivariate SGs to variables that have spatially homogeneous cross-covariance, which is rare in practice.

The application of SGs to large ensembles is relatively new. A SG requires a complex multivariate GST model for its simulations to accurately represent the ensemble distribution of multiple variables. Although developing such a model is a difficult task, the advantages of a SG are such that attempts are valuable. The multivariate GST model underlying the proposed SG can capture complex structures in mean, trend, variance and temporal correlation over

space and spatial correlation over latitudes. However, there are limitations that require model improvements, namely: increased flexibility in modeling tails (e.g., non-Gaussian assumptions), increased flexibility in modeling polar regions and a model specification that can capture spatially heterogeneous cross-correlation structures between variables.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Funding Funding was provided by Engineering and Physical Sciences Research Council (GB) (Grant No. EP/L015358/1).

[Received October 2018. Accepted January 2019. Published Online February 2019.]

APPENDIX A. SPECTRAL MASS FUNCTION SUM

Proof. From (5), the trace of $\mathbf{R}_{m,m,v,v}$ is

$$\begin{aligned} \text{trace}(\mathbf{R}_{m,m,v,v}) &= \text{trace}(\mathbf{W}^{-1} \text{diag}(\mathbf{W}\mathbf{r}_{m,m,v,v})\mathbf{W}) \\ &= \text{trace}(\text{diag}(\mathbf{W}\mathbf{r}_{m,m,v,v})\mathbf{W}\mathbf{W}^{-1}) \\ &= \text{trace}(\text{diag}(\mathbf{W}\mathbf{r}_{m,m,v,v})) \\ &= \sum_{c \in \mathbb{Z}_L} f_{m,v}(c). \end{aligned}$$

Since $\mathbf{R}_{m,m,v,v}$ is a correlation matrix, the trace is also L . □

APPENDIX B. CROSS-SPECTRAL MASS FUNCTION

Proof. The diagonal vector AR model of order one in Sect. 3.7 has the following vector MA model of order m representation

$$\tilde{\mathbf{H}}_{r,t}[c, m] = \sum_{i=0}^m \left(\prod_{j=i+1}^m \Psi_{c,j} \right) \boldsymbol{\epsilon}_{r,t}[c, i], \tag{8}$$

where the product from $m + 1$ to m is defined to be one. Therefore, the element in the row v_1 and column v_2 of

$$\begin{aligned} E(\tilde{\mathbf{H}}_{r,t}[c, m_1] \tilde{\mathbf{H}}_{r,t}^H[c, m_2]) &= \sum_{i=1}^{m_1 \wedge m_2} \left(\prod_{j=i+1}^{m_1} \Psi_{c,j} \right) \boldsymbol{\Xi}_{c,i} \left(\prod_{j=i+1}^{m_2} \Psi_{c,j} \right) \\ &\quad + \left(\prod_{j=1}^{m_1} \Psi_{c,j} \right) \boldsymbol{\Xi}_c \left(\prod_{j=1}^{m_2} \Psi_{c,j} \right) \end{aligned}$$

is

$$\begin{aligned}
 f_{m_1, m_2, v_1, v_2}(c) &= \sum_{i=1}^{m_1 \wedge m_2} \left(\prod_{j=i+1}^{m_1} \psi_{c, j, v_1} \right) \mathbb{E}_{c, i}[v_1, v_2] \left(\prod_{j=i+1}^{m_2} \psi_{c, j, v_2} \right) \\
 &\quad + \left(\prod_{j=1}^{m_1} \psi_{c, j, v_1} \right) \mathbb{E}_c[v_1, v_2] \left(\prod_{j=1}^{m_2} \psi_{c, j, v_2} \right) \\
 &= \mathbb{E}_c[v_1, v_2] \sum_{i=0}^{m_1 \wedge m_2} \left(\prod_{j=i+1}^{m_1} \psi_{c, j, v_1} \right) \left(\prod_{j=i+1}^{m_2} \psi_{c, j, v_2} \right) \\
 &\quad - \mathbb{E}_c[v_1, v_2] \sum_{i=0}^{(m_1 \wedge m_2)-1} \left(\prod_{j=i+1}^{m_1} \psi_{c, j, v_1} \right) \left(\prod_{j=i+1}^{m_2} \psi_{c, j, v_2} \right) \\
 &= \mathbb{E}_c[v_1, v_2] \left(\prod_{j=(m_1 \wedge m_2)+1}^{m_1} \psi_{c, j, v_1} \right) \left(\prod_{j=(m_1 \wedge m_2)+1}^{m_2} \psi_{c, j, v_2} \right) \quad \square
 \end{aligned}$$

REFERENCES

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. et al. (2016), ‘Evaluating lossy data compression on climate simulation data within a large ensemble’, *Geoscientific Model Development* **9**(12), 4381.

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 23rd international symposium on High-performance parallel and distributed computing’, ACM HPDC ’14, pp. 203–214.

Box, G. E. P. and Cox, D. R. (1964), ‘An analysis of transformations’, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 211–252.

Branstator, G. and Teng, H. (2010), ‘Two limits of initial-value decadal predictability in a cpcm’, *Journal of Climate* **23**(23), 6292–6311.

Castruccio, S. (2016), ‘Assessing the spatio-temporal structure of annual and seasonal surface temperature for cmip5 and reanalysis’, *Spatial Statistics* **18**, 179–193.

Castruccio, S. and Genton, M. (2014), ‘Beyond axial symmetry: An improved class of models for global data’, *Stat* **3**, 48–55.

Castruccio, S. and Genton, M. (2018), ‘Principles for inference on big spatio-temporal data from climate models’, *Statistics and Probability Letters* **136**, 92–96.

Castruccio, S. and Genton, M. G. (2016), ‘Compressing an ensemble with statistical models: an algorithm for global 3d spatio-temporal temperature’, *Technometrics* **58**(3), 319–328.

Castruccio, S., Genton, M. and Sun, Y. (2019), ‘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, read before the Royal Statistical Society.

Castruccio, S. and Guinness, J. (2017), ‘An evolutionary spectrum approach to incorporate large-scale geographical descriptors on global processes’, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **66**(2), 329–344.

Castruccio, S., Stein, M. L. et al. (2013), ‘Global space–time models for climate ensembles’, *The Annals of Applied Statistics* **7**(3), 1593–1611.

- Collins, M. (2002), 'Climate predictability on interannual to decadal time scales: the initial value problem', *Climate Dynamics* **19**(8), 671–692.
- Collins, M. and Allen, M. R. (2002), 'Assessing the relative roles of initial and boundary conditions in interannual to decadal climate predictability', *Journal of Climate* **15**(21), 3104–3109.
- Davis, P. J. (2012), *Circulant matrices*, American Mathematical Soc.
- Edwards, M., Castruccio, S. and Hammerling, D. (2018), 'Marginally parametrized spatio-temporal models and stepwise maximum likelihood estimation', [arXiv:1806.11388](https://arxiv.org/abs/1806.11388).
- Friedman, J., Hastie, T. and Tibshirani, R. (2001), *The Elements of Statistical Learning*, Springer.
- Golub, G. H. and Van Loan, C. F. (2012), *Matrix Computations*, Vol. 3, JHU Press.
- Guinness, J. and Hammerling, D. (2018), 'Compression and conditional emulation of climate model output', *Journal of the American Statistical Association* **113**(521), 56–67.
- Hardy, Y. and Steeb, W.-H. (2010), 'Vec-operator, kronecker product and entanglement', *International Journal of Algebra and Computation* **20**(01), 71–76.
- Hitzchenko, M. and Stein, M. L. (2012), 'Some theory for anisotropic processes on the sphere', *Statistical Methodology* **9**, 211–227.
- Hurrell, J. W., Holland, M. M., Gent, P. R., Ghan, S., Kay, J. E., Kushner, P. J., Lamarque, J.-F., Large, W. G., Lawrence, D., Lindsay, K. et al. (2013), 'The community earth system model: a framework for collaborative research', *Bulletin of the American Meteorological Society* **94**(9), 1339–1360.
- Jeong, J., Castruccio, S., Crippa, P., Genton, M. G. et al. (2018), 'Reducing storage of global wind ensembles with stochastic generators', *The Annals of Applied Statistics* **12**(1), 490–509.
- Jeong, J., Yan, Y., Castruccio, S. and Genton, M. (2019), 'A stochastic generator of global monthly wind energy with tukey g-and-h autoregressive processes', *Statistica Sinica* . in press.
- Jones, R. H. (1963), 'Stochastic processes on a sphere', *The Annals of Mathematical Statistics* **34**(1), 213–218.
- Jun, M. (2011), 'Nonstationary cross-covariance models for multivariate processes on a globe', *Scandinavian Journal of Statistics* **38**, 726–747.
- Jun, M. and Stein, M. (2007), 'An approach to producing space-time covariance functions on spheres', *Technometrics* **49**(4), 468–479.
- (2008), 'Nonstationary covariance models for global data', *Annals of Applied Statistics* **2**, 1271–1289.
- Kay, J., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., Arblaster, J., Bates, S., Danabasoglu, G., Edwards, J. et al. (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.
- Lütkepohl, H. (2005), *New introduction to multiple time series analysis*, Springer Science & Business Media.
- Meehl, G. A., Moss, R., Taylor, K. E., Eyring, V., Stouffer, R. J., Bony, S. and Stevens, B. (2014), 'Climate model intercomparisons: preparing for the next phase', *Eos, Transactions American Geophysical Union* **95**(9), 77–78.
- Moss, R., Babiker, W., Brinkman, S., Calvo, E., Carter, T., Edmonds, J., Elgizouli, I., Emori, S., Erda, L., Hibbard, K. et al. (2008), 'Towards new scenarios for the analysis of emissions: Climate change, impacts and response strategies'.
- Pachauri, R. K., Allen, M. R., Barros, V. R., Broome, J., Cramer, W., Christ, R., Church, J. A., Clarke, L., Dahe, Q., Dasgupta, P. et al. (2014), *Climate change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change*, IPCC.
- Patterson, H. D. and Thompson, R. (1971), 'Recovery of inter-block information when block sizes are unequal', *Biometrika* **58**(3), 545–554.
- Paul, K., Mickelson, S., Dennis, J. M., Xu, H. and Brown, D. (2015), Light-weight parallel python tools for earth system modeling workflows, in 'Big Data (Big Data), 2015 IEEE International Conference on', IEEE, pp. 1985–1994.
- Porcu, E., Castruccio, S., Alegria, A. and Crippa, P. (2019), 'Axially symmetric models for global data: a journey between geostatistics and stochastic generators', *Environmetrics* . in press.
- Strand, G. and Baker, A. (2018), Private Communication.

Washington, W. M. and Parkinson, C. (2005), *Introduction to three-dimensional climate modeling*, University Science Books.

Whittle, P. (1954), 'On stationary processes in the plane', *Biometrika* pp. 434–449.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.