Supplementary materials for this article are available at https://doi.org/10.1007/s13253-022-00518-x.

Saving Storage in Climate Ensembles: A Model-Based Stochastic Approach

Huang HUANG, Stefano CASTRUCCIO, Allison H. BAKER, and Marc G. GENTON

While climate models are an invaluable tool for increasing our understanding and therefore, the predictability of the Earth's system for decades, their increase in complexity and resolution has put a considerable, growing strain on the computational resources of research centers and institutions worldwide. The statistics community has a long history of developing stochastic models as a means to save computational time, but the emergence of storage as an additional cost for climate investigations has prompted a reformulation of the aim of statistical models in model-based environmental science. Can stochastic approximations be useful as a mechanism for saving both computational time and storage? We focus on a collection of simulations from a climate model and propose several statistical models of increasing complexity. By analyzing and discussing the associated costs for each model, we demonstrate how computation and storage are closely intertwined, and how a statistical model of increasing complexity is justified only to the extent that information at a fine spatial and/or temporal scale is sought to be preserved.

Supplementary materials accompanying this paper appear online.

Key Words: Climate model; Computational cost; Global model; Space-time model; Stochastic generator; Storage cost.

1. INTRODUCTION

While the environment has been studied using quantitative approaches for centuries, the use of physical models to drive scientific progress has been marginal for all but the last few decades owing to the lack of appropriate tools to solve complex equations. However, the exponential increase in computational power and availability has fueled the unprecedented

H. Huang · M. G. Genton (✉) Statistics Program, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia
(E-mail: *marc.genton@kaust.edu.sa*).
S. Castruccio, Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN 46556, USA.
A. H. Baker, Computational and Information Systems Lab, National Center for Atmospheric Research, Boulder, CO 80305, USA.

Journal of Agricultural, Biological, and Environmental Statistics, Volume 28, Number 2, Pages 324–344
https://doi.org/10.1007/s13253-022-00518-x

development of numerical models for the Earth's system. Climate models provide synthetic representations of an Earth system (or parts thereof) through a collection of equations (partial differential equations, PDEs) representing fundamental laws of physics, such as motion and conservation of momentum and energy. These models have been instrumental in developing our current understanding and hence predictability of the Earth's climate, so much that they are now used not just by scientists, but also by governmental bodies such as the Intergovernmental Panel on Climate Change (IPCC) to deliver periodic assessment reports (AR) with guidelines for policymakers (IPCC 2021).

While climate models have become increasingly more complex in their representations of the Earth's system, they are imperfect because they fail to accurately represent all relevant physical processes. For example, climate simulations from the Coupled Model Intercomparison Phase 6 (CMIP6, Eyring et al. (2016)), the reference simulations for the IPCC AR6 (IPCC 2021), are solved on a spatial scale of the order of tens of kilometers, a resolution too coarse to capture convective boundary layer processes responsible for cloud formation. Therefore, a single climate simulation is insufficient to properly characterize the Earth's system, and a collection (*ensemble*) of simulations is necessary to assess the sensitivity of the results concerning the parameterization of physical processes, natural (unforced) weather variability, and future emission scenarios.

Generating multiple simulations is a burdensome task, as simulating from modern climate models requires weeks to months on high-performance computers available to only a few research centers and institutions worldwide. The non-negligible computational cost has been long acknowledged, and the statistical community has developed a wide range of methods to provide fast stochastic approximations of climate models. At the core of these *emulators* is the idea that a statistical model could be trained with a small number of simulations and then be used to approximate the simulations for unexplored input values (instead of running full climate models). Emulators have been developed for decades, and the standard framework has focused on Gaussian processes in parameter space (Sacks et al. 1989; Kennedy and O'Hagan 2001; Oakley and O'Hagan 2004). Through the years, more articulated models have been proposed to incorporate data resolved in space (Chang et al. 2014), time, space *and* time (Mak et al. 2018), as well as multivariate (Overstall and Woods 2016) and non-Gaussian (Chang et al. 2016) models.

At the core of this work is the acknowledgment that the cost of climate models is not to be evaluated solely in terms of computations. Indeed, over the last few years, storage has become increasingly relevant as a significant limitation of model-based climate science (Baker et al. 2014, 2016). While increases in computational power have enabled, for example, finer resolutions, larger ensembles, and longer simulations, storage technologies have, at least up to the present date, not evolved as fast as computational power. Unless action is taken soon, climate scientists will face hard choices on what data to save, negatively affecting their scientific objectives. For example, here we consider the situation at the National Center of Atmospheric Research (NCAR), one of the prominent research centers in the United States for Earth Systems science. Currently, NCAR's costs for providing and maintaining data, including the costs of hardware, power, staff, and software licenses, are approximately $45 per Terabyte (TB) per year, which translates in costs of hundreds of thousands to millions of dollars per project given that modern simulations often require more

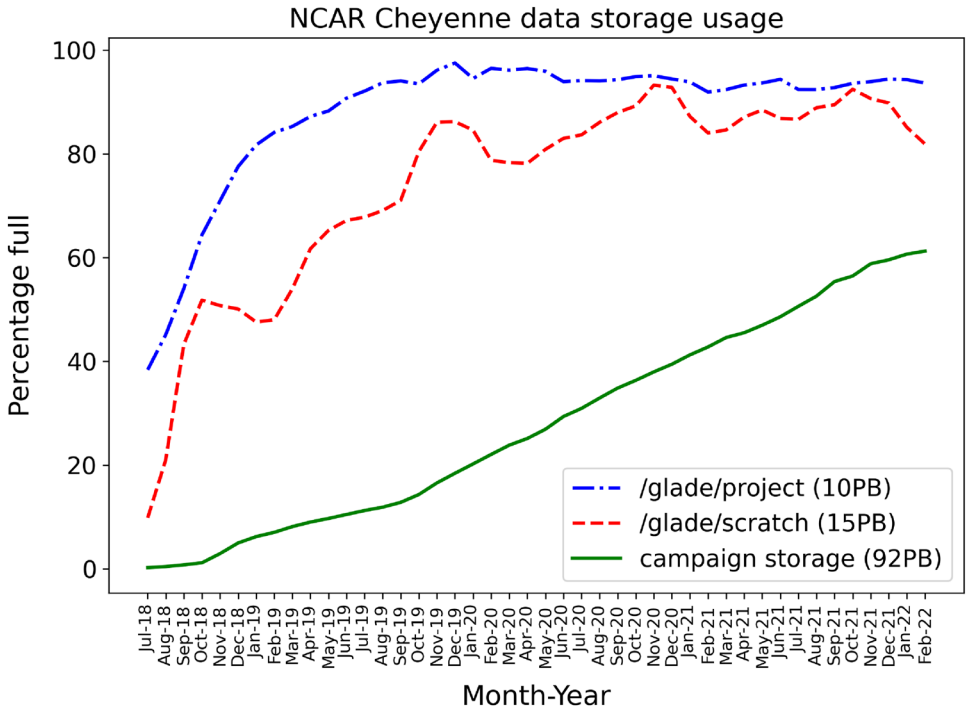

Figure 1. Data storage options and percentage growth at NCAR for its Cheyenne system (Computational and Information 2017). The GLADE file system, which is divided into /project and /scratch collections, is a high-performance (i.e., expensive) shared file system with storage resources that are essentially fixed (due to cost). Campaign storage is a less-expensive resource for medium-term storage of project data that is expandable. The capacity of each space is in parenthesis in the legend.

than 1 Petabyte (PB). Most importantly, while storage demands are continually increasing, available storage resources at institutions such as NCAR are limited due to financial considerations. Although storage for simulations has increased at NCAR to the current capacity of 92 PB to accommodate the growing demand (see Fig. 1), the usable capacity is currently planned to plateau at approximately 100 to 120 PB within the next few years, after which the budget share from storage will not increase and will be used to maintain the current space (i.e., replacing disks that have reached the end of life). The limitation of storage space has long been acknowledged in the climate community, and some *ad hoc* solutions ranging from single-precision storage to subsampling in space, time, or both have been put forward. Compression algorithms have been recently used as means to decrease the storage burden, with solutions spanning from JPEG compression (Woodring et al. 2011; Hübbe et al. 2013) to fpzip and ISABELA (Baker et al. 2014), with a general consensus emerging that no algorithm would be ideal across all variables and levels of temporal aggregation (Baker et al. 2017). More recent studies have acknowledged the need to account for spatial and temporal proximity in order to maximize the compression rate (Bicer et al. 2013; Poppick et al. 2020), and using only the number of bits in the number format which carry scientific information (Kloewer et al. 2021).

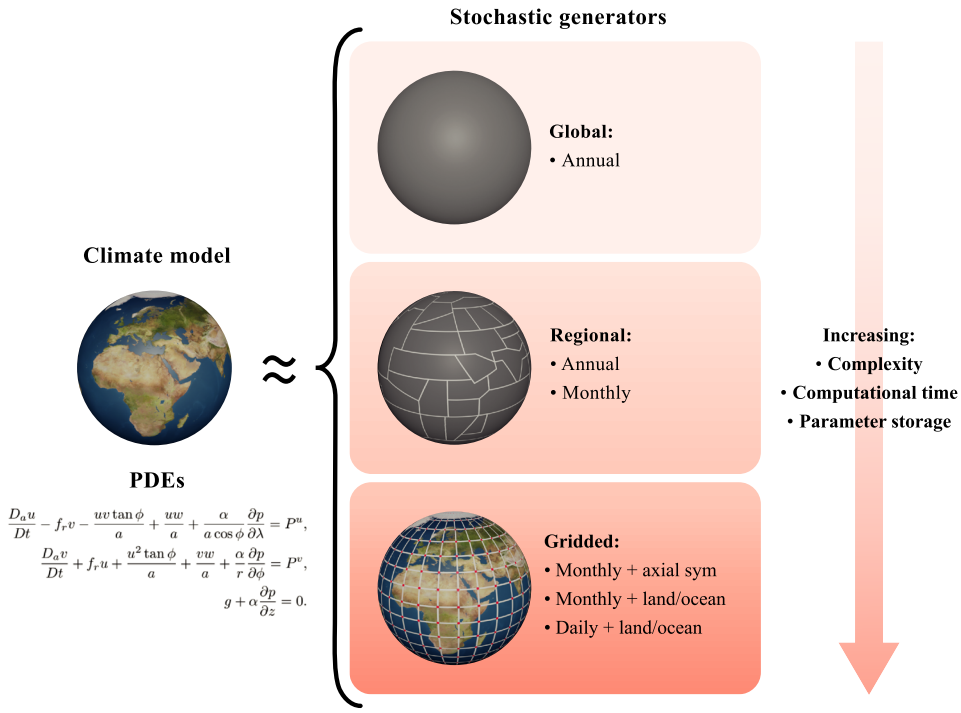

Figure 2. Illustration of stochastic generators of increasing complexity to approximate a climate model output.

From the viewpoint of the statistical community, storage as a cost for climate simulations leads us to question and reformulate the foundational assumption of statistical models as a tool designed to only save computational time. Indeed, if a statistical model is to be used as a means to save space, then an emulator could be considered similar in purpose to a compression algorithm. Formally, a statistics-based compression can be obtained as follows. If a numerical simulation is denoted by $\mathbf{Y}$, and we partition it into a compressed part and a retained part $\mathbf{Y} = (\mathbf{Y}_{\text{compress}}, \mathbf{Y}_{\text{store}})$, then we can assume that

$$\mathbf{Y}_{\text{compress}} \mid \mathbf{Y}_{\text{store}} \sim \mathcal{F}(\boldsymbol{\theta}),$$

for some probability distribution $\mathcal{F}$ controlled by a parameter vector $\boldsymbol{\theta}$. A *conditional* approach assumes that $\mathbf{Y}_{\text{store}} \neq \emptyset$ and is predicated on storing information that is challenging to model and conditionally modelling the rest (Guinness and Hammerling 2018). In this work, we focus on an *unconditional* approach, which assumes that $\mathbf{Y}_{\text{store}} = \emptyset$, so that the only information retained from the original climate model data is the statistical parameters, $\boldsymbol{\theta}$ (Castruccio et al. 2019; Hu and Castruccio 2021). Two convenient features of unconditional compression dictate this choice: 1) it achieves more competitive compression rates because it aims at storing just the statistical model parameters instead of the data; and 2) it is methodologically closer to an emulator because it aims at producing a surrogate simulation that is similar to the original data, rather than recovering the original simulation.

In this work, we aim at addressing the topic of stochastic approximations of climate models from the viewpoint of both storage and computational time by focusing on an ensemble of simulations and proposing a range of statistical models of increasing complexity, able to capture the temporal and spatial variability at increasingly high spatial and temporal resolutions (see Fig. 2 for a schematic illustration of the proposed models). This increase in complexity comes at the cost of an increased parameter space and, hence, the computational time for inference. As such, we will show that stochastic approximations can bring substantial storage savings at various spatial and temporal scales as long as the proposed model is able to capture the complex dependencies implied by climate data.

The manuscript is structured as follows. Section 2 presents the data used in this work, and Sect. 3 demonstrates a model for global data. Section 4 describes a regional model at the annual and monthly resolutions, and Sect. 5 provides a model for monthly and daily native grid resolutions. Section 6 compares the different proposed models in terms of storage and computing costs. Finally, Sect. 7 provides a general perspective on the statistical models for the next generation of climate models.

2. DATA

For the experiments in this work, we used data from a publicly available Community Earth System Model (CESM^{TM}, Hurrell et al. (2013)) project: the CESM Large Ensemble (LENS; Kay et al. 2015). This popular collection of climate simulations was created to study the internal climate variability (and climate change) by isolating the effect of unforced variability (i.e., fluctuations dictated exclusively by weather conditions instead of perturbed physics or future scenarios). The CESM-LENS includes a set of 40 ensemble runs for the period 1920 to 2100 using a fully coupled version of CESM at approximately 1° latitude/longitude resolution, comprising 240 TB of data. In order to have a training set which is as uniform as possible, out of the 40 runs we only consider the 35 performed at NCAR and discard the 5 performed by the University of Toronto as they showed a faster increase in global mean temperatures. The number of simulations is considerably higher than a typical ensemble, which usually comprises of only a few runs per scenario. However, since our aim is to present and validate a general framework, this ensemble represents an ideal benchmark, as a small number of members are used in the training set, and the uncertainty of our statistical surrogate can be compared with the 'ground truth' of the remaining ensemble members.

The 35 simulations begin in 1920 and use historical forcing through 2005. The ensemble spread is generated using slight round-off level differences in the initial atmospheric temperature field. Representative Concentration Pathway (RCP) 8.5 (van Vuuren et al. 2011) forcing is used beginning in 2006, a scenario that reflects near-past and future climate change by assuming a strong increase in emissions (so that at the end of the century, the radiative forcing would be 8.5 Wm^{-2}).

The CESM-LENS data project is ideal for our evaluation because of its climate ensemble, its struggles with storage limitations (Baker et al. 2016), and its availability to the broader climate community. Further, these data have been used in several previous experiments to

assess the influence of data compression algorithms on climate models (Baker et al. 2016; Poppick et al. 2020).

We consider the surface air temperature from the atmospheric component of the model at different temporal and spatial resolutions, depending on the statistical models applied in the following sections. The model's equations are solved on a regular global grid with $M = 192$ latitudes and $N = 288$ longitudes, and we consider data aggregated at daily, monthly, and annual scales. Regional data are obtained by performing a weighted average of each grid point with weights proportional to surface area across the regions (region boundaries are given in Figure S1 in the Supplementary Material), and global data are obtained similarly as the weighted average across the entire globe.

Throughout this work, the temperature for realization r is denoted as $Y_r(t)$, which could be a scalar or vector ($\mathbf{Y}_r(t)$) depending on the level of spatial aggregation, and t represents years, months or days depending on the level of temporal aggregation. Similarly, the model input is expressed as $X(t)$, the radiative forcing that changes the Earth's energy balance under RCP 8.5 and was downloaded from the Potsdam RCP scenario data group (http://www.pik-potsdam.de/~mmalte/rcps).

3. GLOBAL STOCHASTIC GENERATORS

In this section, we consider the simplest setting, where the temperature is aggregated globally and annually, so that the analysis is well approximated by a Gaussian time series. Previous work (Castruccio et al. 2014) has shown that temperature at time t can be modeled as dependent on the past trajectory of radiative forcing through an infinitely distributed lag model (Judge et al. 1980). This approach avoids a causality violation (i.e., so that future forcing would not influence the present temperature). The model is written assuming that the temperature $Y_r(t)$ for realization r at year $(2005+t)$, $t = 1, \ldots, T = 95$ is as follows:

$$Y_r(t) = \beta_0 + \beta_1 X(t) + \beta_2 (1-\rho) \sum_{i=1}^{\infty} \rho^{i-1} X(t-i) + \epsilon_r(t), \tag{1}$$

where $\epsilon_r(t)$ is the residual temporal variability, assumed to be an autoregressive process of order one: $\epsilon_r(t) = \phi \epsilon_r(t-1) + \sigma \nu_r(t)$, where $\epsilon_r(0)$ is zero for notational convenience and $\nu_r(t)$ is a standard Gaussian white noise. There could be in principle temporal heteroskedasticity, and $\nu_r(t)$ could be generalized to account for that. However, as far as the detrended residuals of (1) for temperature are concerned, we did not find any clear pattern in the change in temporal variability across years, as indicated by the 10 years rolling window of the standard deviation of the said residuals in Figure S2. While the global mean temperature does not present evidence of a changing variance in time, this is very likely going to be the case for local and possibly regional temperatures (as well as other variables). As such, a more flexible model could be formulated including temporally varying variances, with different degrees of variability in space (e.g., equatorial regions are expected to show an overall more homogeneous behavior than mid-latitude regions).

Heuristically, the model assumes an intercept β_0, a linear contribution of the present forcing through β_1 and another linear contribution of the past forcing through β_2. Past

forcing has an exponentially decreasing contribution function of the lag from the present time, and the decay is controlled by a parameter $\rho \in (0, 1)$. The term $1 - \rho$ is used for normalization purposes.

The model comprises of a total of six parameters: β_0, β_1, β_2, ρ, ϕ, and σ, and inference is straightforward as, conditional on ρ, (1) is a linear model that can be solved by maximizing the profile likelihood. Once inference is performed, realizations of the statistical model can be generated quickly and efficiently. For consistency with the number of runs in the CESM-LENS dataset, we also generated 35 realizations from the statistical model.

To assess the goodness of fit, we used two metrics, I_{fit} and I_{UQ}, to evaluate the performance of the statistical model. They characterize the lack of fit and variability (uncertainty quantification) of the training data against surrogate simulations from the model:

$$I_{\text{fit}} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T}\{Y_r(t) - \hat{Y}(t)\}^2}{\frac{R}{R-1}\sum_{r=1}^{R}\sum_{t=1}^{T}\{Y_r(t) - \bar{Y}(t)\}^2}, \quad I_{\text{UQ}} = \frac{\text{central\_region\_area}\{\hat{Y}_1(t), \ldots, \hat{Y}_R(t)\}}{\text{central\_region\_area}\{Y_1(t), \ldots, Y_R(t)\}},$$

where $\bar{Y}(t)$ denotes the average of all the available R runs, $\hat{Y}(t)$ is the fitted mean value, and $\hat{Y}_r(t)$ is a generated realization from the fitted statistical model. Given the temporal correlation, the index I_{fit} resembles a lack of fit ratio, albeit a formal test cannot be performed given the presence of temporal dependence. Heuristically, however, the same interpretation stands: if the index is close to 1, then the fitted mean value from the statistical model is as good as the ensemble average, hence representing a good fit (Castruccio et al. 2014). The index I_{UQ} compares the central region area of the simulated data from the stochastic generator to that of the CESM-LENS data. The central region is spanned by half of the curves with the largest modified band depth (López-Pintado and Romo 2009), a commonly used choice of functional data depths. Functional data depths are extended notions of ranks for functional data so that the functional data can be ordered. The central region area can be viewed as the functional interquartile range (see Sun and Genton (2011) for more details), and therefore, I_{UQ} characterizes whether the variation from the stochastic generator represents the internal variability in the CESM-LENS data well. The index I_{UQ} is close to 1 if the variations are similar.

Figure 3(A) illustrates the proposed metrics from the model applied to an increasing number of CESM-LENS members. As measured by I_{fit}, the point estimate accuracy has a sharp decrease for a small training set and then, plateaus relatively fast, whereas the uncertainty assessed by I_{UQ} is essentially unchanged. We chose $R_{\text{train}} = 6$, which results in $I_{\text{fit}} = 1.09$ and $I_{\text{UQ}} = 0.98$, so large enough to guarantee parameter accuracy but small enough to be a realistic number of simulations in other ensembles, such as the CMIP6. Figure 3(B) compares 35 CESM-LENS simulations and 35 surrogates (shifted by 1°C to clarify the visualization, the same number of realization was chosen for a fair visual comparison) from the stochastic generator inferred by $R_{\text{train}} = 6$ CESM-LENS simulations. The ability of the model to capture both the trend and associated uncertainty of these end-of-century projections is apparent. Storage and computational information are provided in summary Table 1 and are discussed in the context of the other models.

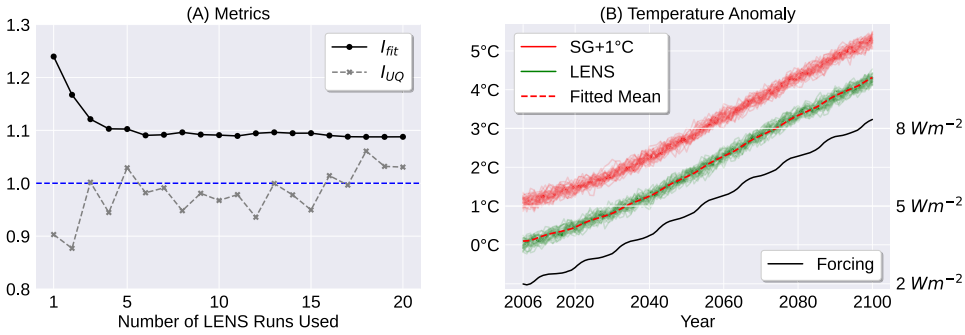

Figure 3. (A) Metrics I_{fit} (solid black) and I_{UQ} (dashed gray) for the global and annual temperature with different numbers of CESM-LENS runs in the training set of Model (1). (B) The 35 CESM-LENS simulations (solid green) and 35 surrogates (solid red) from the stochastic generator (shifted by 1°C upward for clarity) from Model (1) fit with $R_{\text{train}} = 6$ CESM-LENS runs. The dashed red line represents the fitted mean from the statistical model. The radiative forcing is also superimposed (unit measure on the right y axis) to indicate the increasing trend.

4. REGIONALLY AGGREGATED STOCHASTIC GENERATORS

We next consider regional data, where the temperature is regionally aggregated in the same $S = 58$ regions as used in the IPCC AR6 (illustrated in Figure S1 in the Supplementary Material). The temperature for ensemble member r is denoted by $Y_r(s,t)$ at region $s = 1,\ldots,S$ for time t.

4.1. ANNUAL SCALE

For annually aggregated data, we propose an extension of Model (1) to also account for spatial dependence:

$$Y_r(s,t) = \beta_0(s) + \beta_1(s)X(t) + \beta_2(s)\{1-\rho(s)\}\sum_{i=1}^{\infty}\rho^{i-1}(s)X(t-i) + \epsilon_r(s,t), \quad (2)$$

where $\epsilon_r(s,t) = \phi(s)\epsilon_r(s,t-1) + \sigma(s)\nu_r(s,t)$, and $\nu_r(s,t)$ is the standard Gaussian white noise. Regions are dependent in space through the error: if we define $\boldsymbol{\nu}_r(t) = (\nu_r(1,t),\ldots,\nu_r(S,t))^\top$, we assume that $\boldsymbol{\nu}_r(t) \sim \mathcal{N}_S(\mathbf{0},\mathbf{C})$, which is an independent and identically distributed vector across realizations and time. The correlation matrix $\mathbf{C}$ could be parameterized through a spatial model depending on distance, but given the presence of teleconnections (dependence over distant regions due to various atmospheric dynamics), such as the El Niño-Southern Oscillation (ENSO, Philander (1990)), we relied on an unstructured matrix whose only constraints are that it must be symmetric and positive definite.

Given the large parametric space, a single optimization would be computationally infeasible. Instead, to perform inference we propose a two-step approach, where trend and temporal parameters are estimated first and then, the spatial dependence is estimated conditionally on them. Given the large size of the time series, the trend and temporal dependence can generally be estimated with high precision, and previous studies (Castruccio and Stein 2013; Castruccio and Guinness 2017; Castruccio and Genton 2018) showed that error propaga-

tion from the first to the second inferential stage is small, and hence, in this work will be regarded as negligible. In the first step, we misspecified (2) by assuming spatial independence (i.e., $\mathbf{C}$ is instead an identity matrix), which allows the estimation of the trend parameters $\beta_0(s)$, $\beta_1(s)$, $\beta_2(s)$, and $\rho(s)$ and the temporal parameters $\phi(s)$ and $\sigma(s)$ independently for every region by maximizing the profile likelihood, as in the global case in the previous section.

In the second step, conditional on the estimated trend and temporal parameters, we computed the estimated residuals $\hat{v}_r(s,t) = Y_r(s,t) - \hat{Y}(s,t)$, where $\hat{Y}(s,t)$ is the fitted value according to the first step, and we used them to estimate the spatial structure. We estimated the precision matrix nonparametrically by assuming sparsity. If the sample covariance matrix is denoted by $\hat{\mathbf{C}}$, we solved the following minimization (graphical lasso, Friedman et al. (2008)):

$$\hat{\mathbf{C}}_{\text{sparse}}(\lambda) = \arg\min_{\mathbf{C}\geq 0}\{\log\det(\mathbf{C}) + \text{tr}(\mathbf{C}^{-1}\odot\hat{\mathbf{C}}) + \lambda\|\mathbf{P}\odot\mathbf{C}^{-1}\|_1\}, \tag{3}$$

where $\odot$ is the element-wise multiplication, $\mathbf{P}$ denotes a matrix with ones for off-diagonal and zeros for diagonal entries, and λ is a tuning parameter for the induced sparsity penalty. Figure S3(A-B) in the Supplementary Material compares a heatmap of $\hat{\mathbf{C}}^{-1}$ against $\hat{\mathbf{C}}^{-1}_{\text{sparse}}(0.1)$, where $\lambda = 0.1$ was chosen to have approximately 76% zeros. The conditional dependencies for two regions, Central North America (CNA, land) and the North Atlantic Ocean (NAO, ocean), are illustrated in Figure S4 from the sample precision matrix $\hat{\mathbf{C}}^{-1}$ and the sparse approximation $\hat{\mathbf{C}}^{-1}_{\text{sparse}}(0.1)$. Graphical lasso results in both regions having a substantially decreased number of regions with nonzero entries in the precision matrix and being conditionally independent. Indeed, CNA has only 9 nonzero entries, of which by far the largest are the contiguous land regions. NAO is instead related to 15 regions, of which the ones with higher coefficients are the Caribbean and the eastern part of Canada with the Hudson Bay.

The diagnostic metrics I_{fit} and I_{UQ} are shown in Figure S5, and the training set comprising $R_{\text{train}} = 6$ CESM-LENS members achieves stable estimates, with mean (standard deviation) of $I_{\text{fit}} = 1.06(0.02)$ and $I_{\text{UQ}} = 1.03(0.06)$. Once the model is trained, it can produce simulations of spatially dependent regions resembling the original simulations. A marginal comparison of the 35 CESM-LENS simulations against 35 surrogates is shown in Figure S3(C-D), where the proposed Model (2) is shown to be able to capture the higher warming trend over land in CNA, as well as its higher variability.

4.2. Monthly Scale

We now consider monthly data in each region. In this framework, $Y_r(s,t)$ represents the temperature at region s and t months after December 2005. We assume in our statistical model that the radiative forcing $X(t)$ is invariant for different regions and is constant over a calendar year. To account for the interannual temperature trend, we added harmonics to the previous functional form for the mean and a month-specific variance for the error:

$$Y_r(s,t) = \beta_0(s) + \beta_1(s)X(t) + \beta_2(s)\{1-\rho(s)\}\sum_{i=1}^{\infty}\rho^{i-1}(s)X(t-12i) + \sum_{k=1}^{K}\left\{a_k(s)\cos\left(\frac{2\pi tk}{12}\right) + b_k(s)\sin\left(\frac{2\pi tk}{12}\right)\right\} + \epsilon_r(s,t), \tag{4}$$

where $\epsilon_r(s,t) = \phi(s)\epsilon_r(s,t-1) + \sigma(s,t)\nu_r(s,t)$, the parameter $\sigma(s,t)$ is a month-specific standard deviation, and $\boldsymbol{\nu}_r(t) = (\nu_r(1,t),\ldots,\nu_r(S,t))^\top \sim \mathcal{N}_S(\mathbf{0},\mathbf{C})$. As in the annual case, inference was performed in two stages: first the mean parameters $\beta_0(s)$, $\beta_1(s)$, $\beta_2(s)$, $a_k(s)$, $b_k(s)$, where $k = 1,\ldots,K$ and temporal parameters $\phi(s)$ and $\sigma(s,t)$, where $t = 1,\ldots,12\times 95 = 1{,}140$ were estimated by maximizing the profile likelihood. Moreover, K harmonics were considered in the model, and the value of K was determined using the Bayesian information criterion (BIC) shown in Figure S6, where $K = 3$ was chosen. After the mean and temporal parameters were estimated, the spatial dependence was estimated conditionally with a graphical lasso approach similar to the annual case to achieve similar sparsity, for which $\lambda = 0.06$; see Fig. 4(A-B).

As before, we chose the number of runs $R_{\text{train}} = 6$ in the training set using the diagnostics indices, which were evaluated across all regions, and whose mean (standard deviation) is $I_{\text{fit}} = 1.19(0.49)$ and $I_{\text{UQ}} = 1.12(0.20)$, see Figure S7 for a boxplot of these metrics as a function of R. Figure S8 also shows the sample precision matrix along with its sparse counterpart for CNA and NAO, showing approximately the same patterns as in the annual case. To assess the ability of the model to capture the interannual trend, Fig. 4(C-D) compares the 35 CESM-LENS simulations with 35 surrogate runs for CNA and NAO for 2022–2025 and monthly level, and Figure S9 performs the same comparison for annual aggregates. The proposed model is able to capture the different monthly trends across regions, with larger temperature excursions in CNA due to the continental climate compared to slowly varying oceanic patterns in NAO. Despite considerable differences also in the variance, the proposed model appears to be able to capture these features.

5. GRIDDED STOCHASTIC GENERATORS

5.1. MONTHLY SCALE

We now consider monthly temperature at the native model grid scale, so that each ensemble member is simulated globally on a lattice with $M = 192$ latitudes and $N = 288$ longitudes, i.e., $M\times N = 55{,}296$ grid points. We denote with $Y_r(L_m,\ell_n,t)$, $m = 1,\ldots,M$, $n = 1,\ldots,N$, $t = 1,\ldots,T = 1{,}140$, $r = 1,\ldots,R$ the temperature at latitude L_m, longitude ℓ_n, ensemble member r and t months after December, 2005. We assume a location-specific annual and interannual trend along the lines of (1) and (2), with globally spatially dependent innovations:

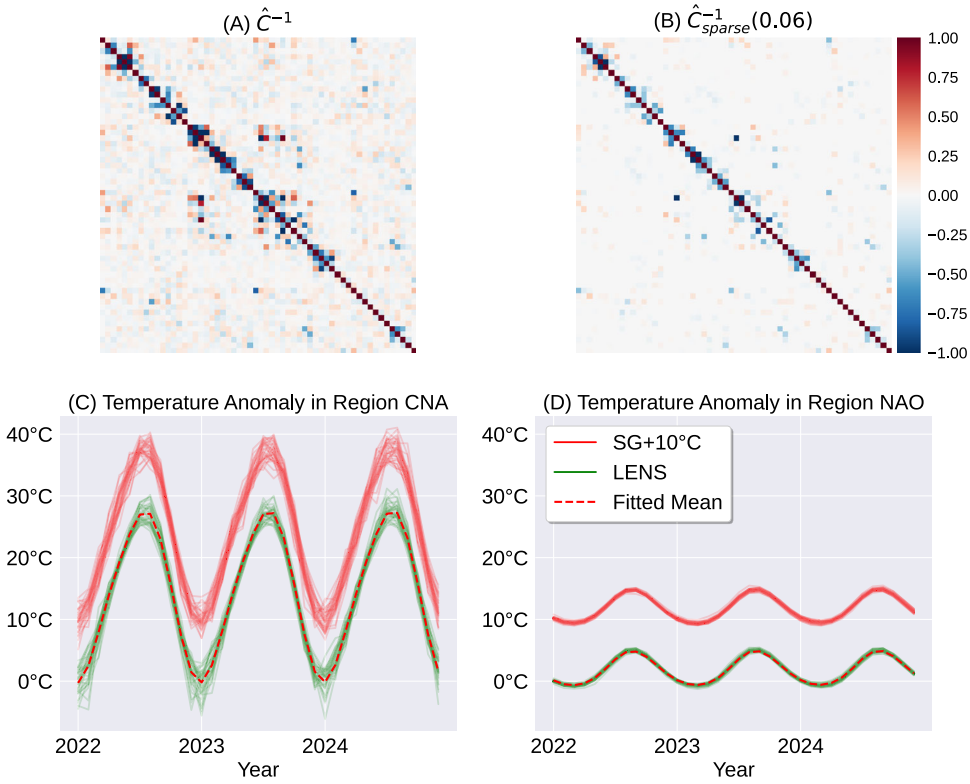

Figure 4. (A) Sample estimate $\hat{\mathbf{C}}^{-1}$ and (B) the sparse estimate $\hat{\mathbf{C}}^{-1}_{\text{sparse}}(0.06)$ from Model (3) with monthly Model (4). (C) and (D): The 35 CESM-LENS simulations (solid green) and 35 surrogates (solid red) from the stochastic generator (shifted by 10°C upwards for clarity) in Central North America (CNA) and North Atlantic Ocean (NAO) regions from Model (4) with (3) fit with $R_{\text{train}} = 6$ CESM-LENS runs. The dashed red line marks the fitted mean from the statistical model.

$$Y_r(L_m, \ell_n, t) = \beta_{m,n,0} + \beta_{m,n,1} X(t) + \beta_{m,n,2}(1 - \rho_{m,n}) \sum_{i=1}^{\infty} \rho_{m,n}^{i-1} X(t - 12i)$$
$$+ \sum_{k=1}^{K} \left\{ a_{m,n,k} \cos\left(\frac{2\pi t k}{12}\right) + b_{m,n,k} \sin\left(\frac{2\pi t k}{12}\right) \right\} + \epsilon_r(L_m, \ell_n, t), \quad (5)$$

where $\epsilon_r(L_m, \ell_n, t) = \phi_{m,n}\epsilon_r(L_m, \ell_n, t-1) + \sigma_{m,n}\eta_r(L_m, \ell_n, t)$ and the vector $\boldsymbol{\eta}_r(t) = (\eta_r(L_1, \ell_1, t), \ldots, \eta_r(L_M, \ell_N, t))^\top \sim \mathcal{N}_{MN}(\mathbf{0}, \boldsymbol{\Sigma})$. The covariance matrix $\boldsymbol{\Sigma}$ needs to account for spatial dependence across the entire globe. Since the observations are provided on a regular grid over the sphere and interpolation is not of primary interest, we do not assume a continuous underlying process, which would imply considerable theoretical challenges (Gneiting 2013). Instead, we rely on a model for a discrete grid, which will be detailed in the inference steps. Consistently with Sect. 4.2, we chose $K = 3$ and used $R_{\text{train}} = 6$ ensemble members.

Similarly to the previous models, inference is performed in a multi-step procedure: 1) trend and time; 2) longitude; and 3) latitude. This approach was shown to be the most

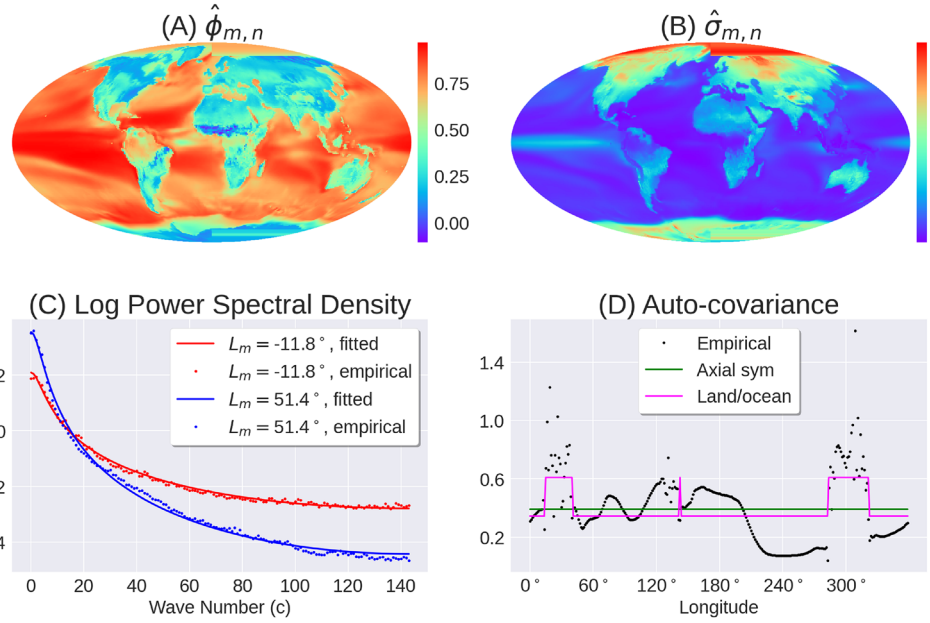

Figure 5. Map of the site-specific autocorrelation $\hat{\phi}_{m,n}$ (A) and standard deviation $\hat{\sigma}_{m,n}$ (B) for the gridded Model (5). (C) Logarithm of fitted and empirical estimates of the power spectral density, $f_{L_m}(c; \psi_m, \alpha_m, \nu_m)$ in (6) for two latitudes $L_m = -11.8°$ and $51.4°$. (D) Comparison of the empirical estimate of $\text{cov}\{\hat{\eta}(L_m, \ell_n), \hat{\eta}(L_m, \ell_{n+1})\}$ at $L_m = -11.8°$ with the estimated one from the axially symmetric and land/ocean models.

efficient choice for gridded global data as it is computationally efficient for large parametric spaces, such as the ones implied by this model, while still retaining asymptotic consistency (Castruccio and Genton 2018; Edwards et al. 2020).

5.1.1. Trend and Time

For each grid point (L_m, ℓ_n), we estimated the mean parameters $\beta_{m,n,j}$, $j = 0, 1, 2$, $a_{m,n,k}, b_{m,n,k}$, $k = 1, \ldots, K$, and the temporal parameters $\rho_{m,n}$ and $\phi_{m,n}$ via profile likelihood. The estimates $\hat{\phi}_{m,n}$ and $\hat{\sigma}_{m,n}$ at all locations are shown in Fig. 5(A-B), where lower auto-correlation and higher variability over land (especially at the poles) is readily apparent.

Once the trend and temporal parameters are estimated, the fitted values are obtained and the innovation estimator $\hat{\boldsymbol{\eta}}_r(t) = (\hat{\eta}_r(L_1, \ell_1, t), \ldots, \hat{\eta}_r(L_M, \ell_N, t))^\top$ is computed, so that the spatial model can be provided. While on a Euclidean domain a standard simplifying assumption is isotropy, i.e., lack of directional dependence, for global data longitude and latitude are expected to have different effects (e.g., variance at mid-latitudes must be higher than at the equator), therefore a more appropriate assumption is that of stationarity across longitudes, or axial symmetry (Jones 1963). While constructive approaches for this class of models have been proposed for a general sampling scheme (Jun and Stein 2008), Castruccio and Stein (2013) have proposed a spectral model for gridded data, which was shown to achieve inference on very large datasets by leveraging on the axially symmetric assumption, and can be extended to multivariate (Edwards et al. 2019) or three-dimensional global data (Castruccio and Genton 2016). In this work, given the similar gridded geometry of the simulated data, we used this approach. Since the spatial structure is assumed to be

independent and identically distributed in time and realization, for simplicity of notation we omit the dependence from these two indices.

5.1.2. Longitude and Latitude

For each latitude L_m, we assumed $\eta(L_m, \ell_n)$ to be stationary in longitude and its spatial dependence was expressed through the power spectral density at wave numbers $c = 0, \ldots, N-1$ as:

$$f(c; \psi_m, \alpha_m, \nu_m) = \sum_{n=0}^{N-1} \exp(-2\pi i c n/N) \mathrm{cov}\{\hat{\eta}(L_m, \ell_1), \hat{\eta}(L_m, \ell_{n+1})\} = \frac{\psi_m}{\left\{\alpha_m^2 + 4\sin^2(c\pi/N)\right\}^{\nu_m + 1/2}}, \tag{6}$$

where $i = \sqrt{-1}$, the parameter $\psi_m > 0$ controls the overall variation, the inverse of $\alpha_m > 0$ determines the range, and $\nu_m > 0$ describes the rate of decay as the wave numbers increase. The MLEs of ψ_m, α_m, and ν_m are obtained independently at each latitude from $\hat{\eta}(L_m, \ell_n)$. Figure S10 depicts $\hat{\psi}_m$, $\hat{\alpha}_m$, and $\hat{\nu}_m$ for each latitude, and Fig. 5(C) shows how the aforementioned functional shape is able to capture the spectral density behavior by comparing the fitted and empirical estimates for two selected latitudes.

Conditionally on the longitudinal parameters, we then consider different latitudinal bands and model their dependence through the coherence of $\hat{\eta}(L_m, \ell_n)$, i.e., its correlation in the spectral domain. If we denote by $f_{L_m, L_{m'}}(c) = \sum_{n=0}^{N-1} \exp(-2\pi i c n/N) \mathrm{cov}\{\hat{\eta}(L_m, \ell_1), \hat{\eta}(L_{m'}, \ell_{n+1})\}$ the cross power spectral density between L_m and $L_{m'}$, then the coherence is modeled as:

$$\rho_{L_m, L_{m'}}(c; \xi, \kappa) = \frac{|f_{L_m, L_{m'}}(c)|}{f(c; \hat{\psi}_m, \hat{\alpha}_m, \hat{\nu}_m) f(c; \hat{\psi}_{m'}, \hat{\alpha}_{m'}, \hat{\nu}_{m'})} = \left[\frac{\xi}{\left\{1 + 4\sin^2(c\pi/N)\right\}^{\kappa}}\right]^{|m-m'|}, \tag{7}$$

where $\xi \in (0, 1)$ determines the overall coherence decay rate with larger latitude differences and $\kappa > 0$ controls the faster coherence decay for larger wave numbers. The computed MLE is $\hat{\xi} = 0.96$ and $\hat{\kappa} = 0.46$. We provide examples of the fitted coherence in Figure S11 in the Supplementary Material.

5.2. Monthly Scale with Land/Ocean Model

While the previous model represents a fast and efficient approach to model axial symmetry, for some variables such as temperature the assumption of stationarity across longitude needs to be relaxed. Indeed, global variables are expected to have different dependence structures when land or ocean is present because ocean temperature is slowly varying in space due to the ocean convection. In order to accommodate this feature, we propose an evolutionary spectrum model which allows for different spectral densities across the two domains. Along the lines in Castruccio and Guinness (2017), this approach assumes that

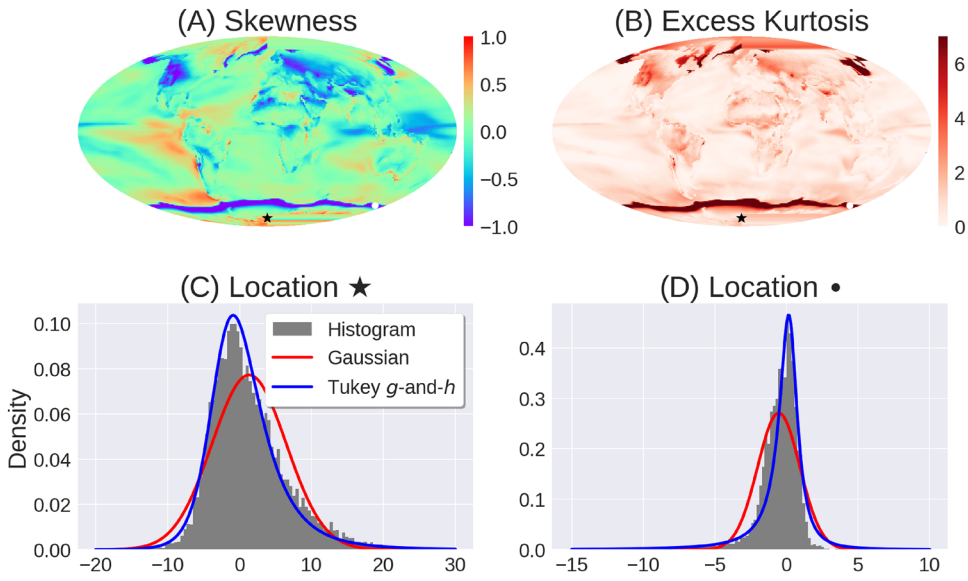

Figure 6. Maps of skewness (A) and excess kurtosis (B) of the estimated detrended residuals $\hat{\epsilon}_r(L_m, \ell_n, t)$. (C-D) Histogram and Gaussian and Tukey g-and-h fit of $\hat{\epsilon}_r(-75.86°, 0°, t)$ and $\hat{\epsilon}_r(-61.73°, 163.75°, t)$, respectively.

the spectral density, f_{L_m,ℓ_n}, is not a constant but varying in longitude:

$$\hat{\eta}_r(L_m, \ell_n, t) = N^{-1/2} \sum_{n=0}^{N-1} \sqrt{f(c, \ell_n; \boldsymbol{\theta}_m^{\text{land}}, \boldsymbol{\theta}_m^{\text{ocean}})} \exp(2\pi i c n/N) \tilde{\eta}_r(L_m, c, t).$$

The longitudinally varying spectral density changes across land and ocean:

$$\begin{aligned} &f(c, \ell_n; \boldsymbol{\theta}_m^{\text{land}}, \boldsymbol{\theta}_m^{\text{ocean}}) \\ &= \left[\sqrt{f(c, \ell_n; \boldsymbol{\theta}_m^{\text{land}})} I_{\text{land}}(L_m, \ell_n) + \sqrt{f(c, \ell_n; \boldsymbol{\theta}_m^{\text{ocean}})}\{1 - I_{\text{land}}(L_m, \ell_n)\}\right]^2, \end{aligned} \quad (8)$$

where $f(c, \ell_n; \boldsymbol{\theta}_m^{\text{j}})$, $\text{j} \in \{\text{land}, \text{ocean}\}$ has the same functional form as (6) with $\boldsymbol{\theta}_m^{\text{j}} = (\psi_m^{\text{j}}, \alpha_m^{\text{j}}, \nu_m^{\text{j}})^\top$, $\text{j} \in \{\text{land}, \text{ocean}\}$. The function $I_{\text{land}}(L_m, \ell_n)$ is the indicator function for whether the location (L_m, ℓ_n) is on land. Figure S12 in the Supplementary Material illustrates all of the estimated parameters for land and for ocean.

To account for the dependence across latitudes, we assume that $\text{corr}\{\tilde{\eta}_r(L_m, c, t), \tilde{\eta}_r(L_{m'}, c', t')\} = \mathbb{I}_{\{c=c', t=t'\}} \rho_{L_m, L_{m'}}(c; \xi, \kappa)$, as defined in (7). We obtained $\hat{\xi} = 0.95$ and $\hat{\kappa} = 0.42$ in this model, and examples of coherence are also provided in Figure S12. Figure 5(D) shows the empirical and fitted covariance between $\hat{\eta}(L_m, \ell_n)$ and $\hat{\eta}(L_m, \ell_{n+1})$ according to the axially symmetric and land/ocean models, and it can be seen how the evolutionary spectrum approach is able to capture the abrupt changes in variability between the two domains.

5.3. DAILY SCALE

Finally, we propose a model for daily data. In this setting, $Y_r(L_m, \ell_n, t), m = 1, \ldots, M$, $n = 1, \ldots, N$, $t = 1, \ldots, T = 34{,}675$, and $r = 1, \ldots, R$ denotes the temperature at latitude L_m, longitude ℓ_n, ensemble r, and t days after December 31, 2005. We propose the same model as in the monthly scale for gridded data (5). However, the Gaussian distribution for the model residuals $\epsilon_r(L_m, \ell_n, t)$ is inadequate for daily resolution. Indeed, the Jarque–Bera test results in 99.89% locations (55,236 out of 55,296) rejecting Gaussianity, and from Fig. 6(A-B) it can be seen how large portions of the world are negatively skewed and there is high excess kurtosis at the poles. In order to account for a more flexible marginal distribution, we propose a trans-Gaussian model with the Tukey g-and-h transformation (Jeong et al. 2019), which allows to control moments of higher orders with two separate parameters. Formally, we assume that $\epsilon_r(L_m, \ell_n, t) = \omega_{m,n}\tau_{g_{m,n},h_{m,n}}(\tilde{\epsilon}_r(L_m, \ell_n, t))$ where

$$\tau_{g_{m,n},h_{m,n}}(z) = \begin{cases} g_{m,n}^{-1}\{\exp(g_{m,n}z) - 1\}\exp(h_{m,n}z^2/2), & \text{if } g_{m,n} \neq 0, \\ z\exp(h_{m,n}z^2/2), & \text{if } g_{m,n} = 0, \end{cases}$$

is the Tukey g-and-h transformation. As before, $\tilde{\epsilon}_r(L_m, \ell_n, t) = \phi_{m,n}\tilde{\epsilon}_r(L_m, \ell_n, t-1) + \sigma_{m,n}\eta_r(L_m, \ell_n, t)$, where $\boldsymbol{\eta}_r(t) = (\eta_r(L_1, \ell_1, t), \ldots, \eta_r(L_M, \ell_N, t))^\top \sim \mathcal{N}_{MN}(\mathbf{0}, \boldsymbol{\Sigma})$ with the same covariance structure as in Sect. 5.2. Consistently with previous sections, we choose $K = 3$ and use $R_{\text{train}} = 6$ ensemble members in the training set. To ease the computational burden implied by the massive amount of data, nearly 2 billions space-time points, inference is performed only on daily data for 2020, 2040, 2060, 2080 and 2100. Figure 6(C-D) depicts the fit of the Tukey g-and-h model for two selected points, along with the best fit from the Gaussian model. The inadequacy of the Gaussian distribution is readily apparent, as the estimated residuals show substantial asymmetry and more generally non-Gaussian higher moments. It is also clear that the proposed model is not particularly suitable for capturing extreme events, but this is to be somewhat expected as its main aim is to capture the entire probability distribution of daily data. We envision two possible solutions to address this:

1. A single model for both the center of the distribution and extremes or other aspects such as threshold exceedances could be formulated.
2. Multiple models could be proposed, each with focus on a different aspects of the distribution.

As of today, the literature on unifying models for extremes, threshold exceedances and the center of the distribution is sparse and far from being complete and operational. As such, we believe that it would be more practical to propose a small set of compression models for different aspects of the distribution rather than a single comprehensive model, especially since the necessary space to store parameters from a few models is still considerably smaller than traditionally compressed data, as will be shown in more details in the next section.

Table 1. Summary of all models used in terms of number of parameters used (#Para), storage required (Storage), inference and simulation time (I. Time, and S. Time, respectively)

Model	#Para	Storage	I. Time	S. Time
Global-annual	6	48 B	0.5 secs	negligible
Regional-annual	2,059	16.09 KB	0.93 secs*	negligible
Regional-monthly	3,045	23.79 KB	5.28 secs*	0.1 secs
Gridded-monthly-axial-sym	664,130	5.07 MB	1.46 mins‡	4.07 mins
Gridded-monthly-land/ocean	664,706	5.07 MB	10.88 mins‡	4.88 mins
Gridded-daily-land/ocean	830,594	6.34 MB	48.07 mins‡	1.94 hours

Inference time is reported by performing the inference of the mean trend in parallel for each region* or for each grid point‡, and in parallel for the spatial dependence along longitudes for different latitudinal bands‡

6. MODEL COMPARISON

The models used in this work (see Fig. 2 for a synoptic view) are summarized and compared in terms of number of parameters, required storage (excluding the space for the programs implementing the statistical model), inference and simulation times in Table 1. The reported times are for execution on a computer with a 52-core Intel Xeon Gold 6230R 2.10GHz CPU, while the mean trend is inferred in parallel for each region or for each grid point and each process uses one core of AMD EPYC 7702 2.00GHz CPU. For the gridded models, the spatial dependence along longitudes is also inferred in parallel for different latitudinal bands. Different computing hardware will inevitably achieve inference with different computational time, but the change will not be to the extent that it would alter our conclusions.

The table has some apparent patterns as the models increase in complexity. Firstly, as we focus on higher spatial and temporal scales, the first column highlights a dramatic increase in parameters, from 6 to 830,594, with a change of approximately three orders of magnitude from global to regional and three more from regional to gridded. This increase is mostly dictated by the presence of region-specific or location-specific trend and temporal parameters. We choose not to reduce the number of parameters by proposing a spatial model or some form of compression for trend and temporal parameters, as in terms of storage the impact of such parameters is minimal. Indeed, from the second column it is apparent how this steep increase in the number of parameters only implies an additional storage of a few megabytes, a negligible amount for present-day computers compared to the approximately 500 GB of storage of the temperature for the full LENS. The use of such little space can be mostly attributed to three features of our proposed model: 1) a simple yet effective parametrization of annual and interannual trends and temporal dependence; 2) a flexible and parsimonious spatial model tailored to the nature of global data; and 3) the ability of our models to reproduce not just one simulation, but the features of an entire ensemble (unconditional compression).

In terms of computational time for inference and simulation, the last two columns of Table 1 also highlight an increase as the model complexity increases, albeit not to the same extent as storage. Global and regional models have an almost negligible computational cost,

as they require less than a few seconds for inference and simulation. Gridded models instead rely on a multi-step approach which requires less than an hour (with full parallelization) for inference and approximately two for simulations for the daily case (due to the very large number of days to be simulated). The computational efficiency of the proposed models can be mostly attributed to: 1) multi-step approaches, which allow parallelization of parts of inference; 2) grid-specific, spectral methods to model the spatial dependence; and 3) the use of multiple simulations in the training set to improve parameter identifiability, see Castruccio and Genton (2018) for a complete discussion.

7. DISCUSSION

While computational cost has long been acknowledged as a limiting factor in model-based climate science, storage is now on the trajectory to become a significant limitation for the sustainable usage of ever more complex climate models. In the authors' view, it is not a matter of **if**, but rather of **when** this topic will need to be systematically addressed with formal methods. Given the long and well-established history of stochastic models for emulation, this paradigm shift represents an unprecedented opportunity for the statistical community to provide critical contributions to the next generation of weather and climate models. This different perspective raises significant challenges in developing new statistical models: Are statistical models, as currently defined, the best way to perform statistics-based compression?

Our results suggest that, as long as the aim is to reproduce the distribution from which the climate simulations are drawn rather than the simulations themselves, the models available to date provide a means to store information from climate simulations at a negligible fraction of the storage cost of the entire ensemble. As such, there is a lot of room for development of a wide range of statistical models without the cost of storing parameters becoming a substantial concern. This work has described a systematic comparison of previous literature under the broad topic of proposing multiple stochastic approximations to the same climate model and has provided evidence that, for a specific ensemble, the classical statistical approaches are indeed suitable compression methods. Future methods could diverge from the traditional emulator literature, especially since the ensemble used in this work was specifically designed to isolate internal variability, and the task of compressing ensembles with different scenarios or physics was avoided. It is, however, conceivable that the proposed framework can be extended to more general ensembles, and a range of models can be designed controlled by the storage quota for the parameters, as variables with more complex behavior such as high-resolution precipitation or wind would require more articulated methods such as latent Gaussian models. Additionally, in order to be widely used and allow to extract valuable scientific information, the proposed framework will have to be extended to multivariate models and capture behavior of multiple physical variables at the same time. Statistical models for multiple spatio-temporal variables are widely acknowledged to be extremely challenging to construct, see Genton and Kleiber (2015) for a general review. As such, in line with the proposed principle of storing only the necessary information and conditionally model the rest, stochastic compression for multivariate models could be designed so that challenging vari-

ables would be stored either in full or partially with some dimensional reduction approach, while variables with simpler behavior could be conditionally stochastically modeled.

The use of stochastic methods to compress climate models is also bound to shape the development of diagnostics tools. In this regard, conditional compression aims to retrieve the original data by storing part of it, and then, aiming at producing an uncompressed dataset which would not be distinguishable from the original data, a form of Turing test for investigations in climate science. The unconditional approach in this work aims instead to produce new surrogate realizations with analogous physical properties, in a similar fashion to Stochastic Weather Generators (SWGs, Richardson (1981)). Unlike SWGs, which have been traditionally developed in the context of time series, compression diagnostics are expected to have a significant spatial component. Thus, previous work has indicated that the image processing literature could provide quantitative metrics to assess image similarities in the context of two-dimensional figures and movies and virtual-reality environments (Castruccio et al. 2019).

The aforementioned points are valid only to the extent that no major technological breakthrough will alter the current relative trends in the cost of storage and computation. While technologies such as DNA storage and molecular memory could revolutionize the current approach to store information and disrupt the balance between cost and computation, their technological development is still at the proof-of-concept stage. In the foreseeable future, environmental science is expected to continue to rely on hard disk drives, solid-state memory, or tapes.

Our results also highlight that inference and simulation require at most hours for models on the ensemble native grid with negligible storage costs. While these results underscore how our models require some degree of computational effort, we argue that in the context of climate simulations, this cost is not a major concern: If a single climate simulation requires multiple weeks on a computing cluster, a few hours to perform inference for a statistical model able to reproduce an entire ensemble with 35 simulations is comparatively very little time that can and should be allocated. Additionally, when inference is performed, the computational cost for surrogate simulations is minimal: They can be produced in a short time on any computer. Furthermore, previous work (Jeong et al. 2018) has also shown how user-friendly interfaces can be developed to allow environmental scientists to generate data without in-depth knowledge of the statistical model.

Although our work has focused on one ensemble and variable, our results suggest that stochastic approximations can be a useful means to preserve not just computations, but also storage, as long as practitioners are willing to accept the notion of approximating a climate model.

ACKNOWLEDGEMENTS

Fig. 2 was produced by Antonio García, scientific illustrator. We thank Dave Hart (NCAR) for the data for Fig. 1, Cecile Hannay (NCAR) for the discussion about technical details of the climate model, and the review team for comments that improved this manuscript. This research was supported by the King Abdullah University of Science and Technology (KAUST).

Declarations

Data Availability The code for this work is available at the GitHub repository: https://github.com/hhuang90/stochastic_emulator. The data are freely available at the Earth Grid System repository.

[Received May 2022. Revised September 2022. Accepted September 2022. Published Online May 2023.]

REFERENCES

Baker AH, Hammerling DM, Mickelson SA, Xu H, Stolpe MB, Naveau P, Sanderson B, Ebert-Uphoff I, Samarasinghe S, De Simone F, Carbone F, Gencarelli CN, Dennis JM, Kay JE, Lindstrom P (2016) Evaluating lossy data compression on climate simulation data within a large ensemble. Geosci Model Dev 9:4381–4403

Baker A H, Xu H, Dennis J M, Levy M N, Nychka D, Mickelson SA, Edwards J, Vertenstein M, 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, pp 203–214. ACM HPDC '14

Baker AH, Xu H, Hammerling DM, Li S, Clyne JP (2017). Toward a multi-method approach: lossy data compression for climate simulation data. High performance computing, pp. 30–42. Springer, Berlin

Bicer T, Yin J, Chiu D, Agrawal G, Schuchardt K (2013). Integrating online compression to accelerate large-scale data analytics applications. Parallel and distributed processing symposium, international, pp 1205–1216

Castruccio S, Genton MG (2016) Compressing an ensemble with statistical models: An algorithm for global 3D spatio-temporal temperature. Technometrics 58(3):319–328

Castruccio S, Genton MG (2018) Principles for statistical inference on big spatio-temporal data from climate models. Stat Probab Lett 136:92–96

Castruccio S, Genton MG, Sun Y (2019). Visualizing spatiotemporal models with virtual reality: From fully immersive environments to applications in stereoscopic view. J R Stat Soc Ser A *182*(2), 379–387

Castruccio S, Guinness J (2017) An evolutionary spectrum approach to incorporate large-scale geographical descriptors on global processes. J Roy Stat Soc Ser C (Appl Stat) 66(2):329–344

Castruccio S, Hu Z, Sanderson B, Karspeck A, Hammerling D (2019) Reproducing internal variability with few ensemble runs. J Clim 32(24):8511–8522

Castruccio S, McInerney DJ, Stein ML, Liu Crouch F, Jacob RL, Moyer EJ (2014) Statistical emulation of climate model projections based on precomputed GCM runs. J Clim 27(5):1829–1844

Castruccio S, Stein ML (2013) Global space-time models for climate ensembles. Ann Appl Stat 7:1593–1611

Chang W, Haran M, Applegate P, Pollard D (2016) Calibrating an ice sheet model using high-dimensional binary spatial data. J Am Stat Assoc 111(513):57–72

Chang W, Haran M, Olson R, Keller K (2014) Fast dimension-reduced climate model calibration and the effect of data aggregation. Ann Appl Stat 8(2):649–673

Computational and Information Systems Laboratory (2017). Cheyenne: SGI ICE XA Cluster

Edwards M, Castruccio S, Hammerling D (2019) A multivariate global spatiotemporal stochastic generator for climate ensembles. J Agric Biol Environ Stat 24:464–483

Edwards M, Castruccio S, Hammerling D (2020) Marginally parameterized spatio-temporal models and stepwise maximum likelihood estimation. Comput Stat Data Anal 151:107018

Eyring V, Bony S, Meehl GA, Senior CA, Stevens B, Stouffer RJ, Taylor KE (2016) Overview of the coupled model intercomparison project phase 6 (CMIP6) experimental design and organization. Geosci Model Dev 9(5):1937–1958

Friedman J, Hastie TR (2008) Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9(3):432–441

Genton MG, Kleiber W (2015) Cross-covariance functions for multivariate geostatistics (with discussion). Stat Sci 30(2):147–163

Gneiting T (2013) Strictly and non-strictly positive definite functions on spheres. Bernoulli 19(4):1327–1349

Guinness J, Hammerling D (2018) Compression and conditional emulation of climate model output. J Am Stat Assoc 113(521):56–67

Hu W, Castruccio S (2021) Approximating the internal variability of bias-corrected global temperature projections with spatial stochastic generators. J Clim 34:8409–8418

Hübbe N, Wegener A, Kunkel JM, Ling Y, Ludwig T (2013) Evaluating lossy compression on climate data. In: Kunkel JM, Ludwig T, Meuer HW (eds) Supercomputing. Springer, Berlin Heidelberg, pp 343–356

Hurrell J, Holland M, Gent P, Ghan S, Kay J, Kushner P, Lamarque J-F, Large W, Lawrence D, Lindsay K, Lipscomb W, Long M, Mahowald N, Marsh D, Neale R, Rasch P, Vavrus S, Vertenstein M, Bader D, Collins W, Hack J, Kiehl J, Marshall S (2013) The community earth system model: a framework for collaborative research. Bull Am Meteor Soc 94:1339–1360

IPCC (2021). IPCC, 2021: summary for policymakers. In: Climate change 2021: the physical science basis. contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change. Cambridge University Press

Iturbide M, Gutiérrez JM, Alves LM, Bedia J, Cerezo-Mota R, Cimadevilla E, Cofiño AS, Di Luca A, Faria SH, Gorodetskaya IV, Hauser M, Herrera S, Hennessy K, Hewitt HT, Jones RG, Krakovska S, Manzanas R, Martínez-Castro D, Narisma GT, Nurhati IS, Pinto I, Seneviratne SI, van den Hurk B, Vera CS (2020) An update of IPCC climate reference regions for subcontinental analysis of climate model data: definition and aggregated datasets. Earth Syst Sci Data 12(4):2959–2970

Jeong J, Castruccio S, Crippa P, Genton MG (2018) Reducing storage of global wind ensembles with stochastic generators. Ann Appl Stat 12(1):490–509

Jeong J, Yan Y, Castruccio S, Genton MG (2019) A stochastic generator of global monthly wind energy with Tukey g-and-h autoregressive processes. Stat Sin 29(3):1105–1126

Jones RH (1963) Stochastic processes on a sphere. Ann Math Stat 34:213–218

Judge G, Griffiths W, Hill E, Lutkepohl H, Lee T-S (1980) The theory and practice of econometrics. Wiley, Hoboken

Jun M, Stein ML (2008) Nonstationary covariance models for global data. Ann Appl Stat 2(4):1271–1289

Kay JE, Deser C, Phillips A, Mai A, Hannay C, Strand G, Arblaster JM, Bates S, Danabasoglu G, Edwards J, Holland 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 Meteor Soc 96(8):1333–1349

Kennedy MC, O'Hagan A (2001) Bayesian calibration of computer models. J R Stat Soc Ser B (Stat Methodol) 63(3):425–464

Kloewer M, Razinger M, Dominguez J, Dueben P, Palmer T (2021). Compressing atmospheric data into its real information content. www.researchsquare.com/article/rs-590601/v1

López-Pintado S, Romo J (2009) On the concept of depth for functional data. J Am Stat Assoc 104(486):718–734

Mak S, Sung C-L, Wang X, Yeh S-T, Chang Y-H, Joseph VR, Yang V, Wu CFJ (2018) An efficient surrogate model for emulation and physics extraction of large eddy simulations. J Am Stat Assoc 113(524):1443–1456

Oakley JE, O'Hagan A (2004) Probabilistic sensitivity analysis of complex models: a Bayesian approach. J R Stat Soc Ser B (Stat Methodol) 66(3):751–769

Overstall AM, Woods DC (2016) Multivariate emulation of computer simulators: model selection and diagnostics with application to a humanitarian relief model. J Roy Stat Soc: Ser C (Appl Stat) 65(4):483–505

Philander SG (1990) El Niño, La Niña, and the southern oscillation. Academic Press, Cambridge

Poppick A, Nardi J, Feldman N, Baker AH, Pinard A, Hammerling DM (2020) A statistical analysis of Lossily compressed climate model data. Comput Geosci 145:104599

Richardson CW (1981) Stochastic simulation of daily precipitation, temperature, and solar radiation. Water Resour Res 17(1):182–190

Sacks J, Welch WJ, Mitchell TJ, Wynn HP (1989) Design and analysis of computer experiments. Stat Sci 4(4):409–423

Sun Y, Genton MG (2011) Functional boxplots. J Comput Graph Stat 20(2):316–334

van Vuuren DP, Jae Edmonds MK, Riahi K, Thomson A, Hibbard K, Hurtt GC, Kram T, Krey V, Lamarque J-F, Masui T, Meinshausen M, Nakicenovic N, Smith SJ, Rose SK (2011). The representative concentration pathways: an overview. Clim Change, 109(5)

Woodring J, Mniszewski SM, Brislawn CM, DeMarle DE, Ahrens JP (2011) Revisiting wavelet compression for large-scale climate data using JPEG2000 and ensuring data precision. In: Rogers D, Silva CT (eds) IEEE symposium on large data analysis and visualization (LDAV). IEEE, pp 31–38