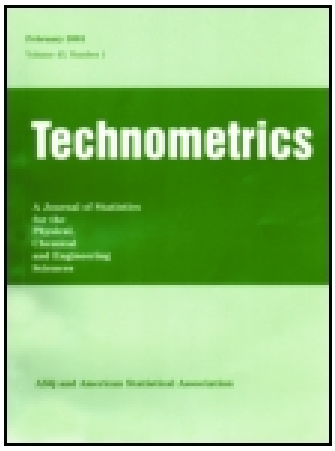

Technometrics

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/utch20

Compressing an Ensemble with Statistical Models: An Algorithm for Global 3D Spatio-Temporal Temperature

Stefano Castruccio[a] & Marc G. Genton[b]

[a] School of Mathematics & Statistics, Newcastle University, Newcastle Upon Tyne, NE1 7RU United Kingdom. E-mail:

[b] CEMSE Division, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia. E-mail:

Accepted author version posted online: 02 Apr 2015.

Compressing an Ensemble with Statistical Models: An Algorithm for Global 3D Spatio-Temporal Temperature

Stefano Castruccio[1] and Marc G. Genton[2]

March 25, 2015

Abstract

One of the main challenges when working with modern climate model ensembles is the increasingly larger size of the data produced, and the consequent difficulty in storing large amounts of spatio-temporally resolved information. Many compression algorithms can be used to mitigate this problem, but since they are designed to compress generic scientific data sets, they do not account for the nature of climate model output and they compress only individual simulations. In this work, we propose a different, statistics-based approach that explicitly accounts for the space-time dependence of the data for annual global three-dimensional temperature fields in an initial condition ensemble. The set of estimated parameters is small (compared to the data size) and can be regarded as a summary of the essential structure of the ensemble output; therefore, it can be used to instantaneously reproduce the temperature fields in an ensemble with a substantial saving in storage and time. The statistical model exploits the gridded geometry of the data and parallelization across processors. It is therefore computationally convenient and allows to fit a non-trivial model to a data set of one billion data points with a covariance matrix comprising of 10^{18} entries.

Key words: big data; distributed computing; space-time statistics; sphere

Short title: Compression of an Ensemble with Statistics

[1]School of Mathematics & Statistics, Newcastle University, Newcastle Upon Tyne, NE1 7RU United Kingdom. E-mail: stefano.castruccio@ncl.ac.uk

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

1 Introduction

One of the main contemporary scientific problems is how climate is changing, what the patterns of local change are and what the social and economic costs of such changes will be (Meehl et al., 2007). Although weather observations from multiple sources and appropriate statistical analyses can be used to answer some of these questions, it is not possible to separate anthropogenic from natural contributions to global warming since they are confounded in observational data. A popular approach is to use climate models, systems of partial differential equations (PDEs) that seek to reproduce the main physical processes of the Earth's climate. Modern climate models are solved on fine spatio-temporal grids in land, ocean, atmosphere, sea-ice and land-ice for tens of physical variables, and an output for a single simulation can require many Tera bytes of space. A collection (ensemble) of multiple runs and climate models such as the Coupled Model Intercomparison Phase 5 (CMIP5) ensemble requires a large effort from multiple institutions (the Earth System Grid Federation) to store, control and coordinate the data access. It is therefore of paramount importance to develop methods for efficiently compressing available climate model output without substantially reducing the geophysical information. Data compression for climate data has been discussed (Woodring et al., 2011; Hübbe et al., 2013; Bicer et al., 2013), as an application of lossless (Lindstrom and Isenburg, 2006; Burtscher and Ratanaworabhan, 2007; Schendel et al., 2012; Gomez and Cappello, 2013) and lossy algorithms (Lakshminarasimhan et al., 2011; Laney et al., 2013) to scientific data. Recently Baker et al. (2014) reviewed some well-known lossless and lossy compression algorithms for climate data and introduced some diagnostics to understand if and to what extent the compressed data set differs from the original climate model output. The diagnostic was performed on aggregating pointwise discrepancy measures over the spatial domain, to assess if the compressed data set was reproducing the actual value of the computer model.

In this work we propose a new approach on data compression. We focus on an initial condition ensemble and we define a statistical model that explicitly accounts for the spatio-temporal dependence of the data and uses its parameters as summary of the geophysical characteristics of the climate models. We further propose some validation criteria from space-time statistics literature to prove that the statistical model can reproduce the spatio-temporal dependence of the original climate model. This approach allows for compressing an entire ensemble and not a single climate model run at much higher rate than traditional algorithms (here we achieve a 50:1 ratio compared to the best performance of 5:1 in Baker et al. (2014)). The proposed approach, however, hinges on the validity of the assumption of the statistical model, so data diagnostics are important to validate the efficiency of the compression and they are thoroughly discussed. Once fitted, the statistical model allows to conditionally simulate climate model runs with different initial conditions. In this regard, the statistical model can be regarded as an emulator of an initial condition ensemble, under the assumption that runs are independent for different initial conditions. This is, to our knowledge, the first time an emulator is used in this context, as it is traditionally used for calibration and sensitivity analysis (Sansó et al., 2008; Sansó and Forest, 2009; Bhat et al., 2012; Drignei et al., 2008; Chang et al., 2015) or scenario extrapolation (Holden and Edwards, 2010; Castruccio and Stein, 2013; Holden et al., 2013; Castruccio et al., 2014). The key difference with traditional emulators is that we do not assume correlation among inputs, as different initial conditions sensibly sampled from the spin-up run generate effectively independent runs.

The model we propose focuses on annual three dimensional global spatio-temporal temperature fields with more than 1 billion data points and fitting a statistical model on such a large data set is a challenging task. In the case of Gaussian processes, the analysis of a space/time data set of size n with a full dependence structure implies storing matrices with $O(n^2)$ elements, which is a daunting task for data sets that are larger than 50,000 data points with current RAM capabilities.

In addition, the likelihood requires $O(n^3)$ flops for Cholesky decomposition and determinant evaluation. Many approaches have been proposed in recent years to overcome these problems (see Sun et al. (2012) and references therein for a complete review). Among the most popular are reducing the matrix size via a low rank approximation, kernel convolution (Higdon, 1998), fixed rank kriging (Cressie and Johannesson, 2008) and predictive processes (Banerjee et al., 2008). The latter approach is computationally efficient but can lead to loss of information when the spatial correlation is moderate or strong (Stein, 2014). Another approach involves sparse approximation of the covariance matrix via tapering (Furrer et al., 2006) or its inverse via Gaussian Markov Random Field approximation (Rue and Held, 2005; Lindgren et al., 2011; Simpson et al., 2012; Xu et al., 2015), but both these methods still imply a loss of information which depends on the taper size or on the degree of Markovian approximation. Another possibility is using composite likelihoods by assuming independence across blocks (Vecchia, 1988; Stein et al., 2004; Eidsvik et al., 2014), but this approach implies a subjective choice of the blocks and does not allow to model dependence at the boundaries of the blocks (therefore still implying loss of information about the data structure). A recent direction of investigation involves finding the maximum likelihood estimator by finding the zeros of an approximation of the score functions, via the Hutchinson estimator of the matrix trace (Anitescu et al., 2012; Stein et al., 2012).

In this work, we circumvent some of the challenges of fitting unstructured spatio-temporal data by exploiting the gridded geometry of the data and proposing an algorithm for likelihood evaluation that balances memory storage, distributed access to memory and synchronization among processors. These features are strongly dependent on the computer's specifics and will be discussed in detail throughout this work. The multi-stage algorithm we propose in this work captures well the patterns in the data both in time and in space, requires less than 48 hours to run and has approximately 27 million parameters, a small amount ($\approx 2\%$) compared to the data size.

The remainder of the paper is organized as follows: Section 2 introduces the dataset, Section 3 describes how multiple runs in an ensemble allows the estimation of the stochastic part without a model for the mean, Section 4 describes the statistical model, presents the diagnostics, and discusses computational challenges, Section 5 shows how the model can be used to simulate runs from the initial condition ensemble, and Section 6 draws some conclusions.

2 The temperature data set

In this work, we focus on CMIP5 (Taylor et al., 2012), a multi-model ensemble that aims to provide a uniform and comparable assessment of climate response under different climate models for the fifth Intergovernmental Panel on Climate Change (IPCC) Assessment Report. In particular, we focus on the National Center for Atmospheric Research (NCAR) Community Climate System Model 4 (CCSM4; Gent et al. (2011)), under a Representative Concentration Pathway 85 scenario (Van Vuuren et al., 2011). Our choice of model and scenario was based on data availability: this ensemble consists of six realizations (runs with different initial conditions), although the analysis we present can be extended to multiple scenarios. The data set consists of projections of yearly temperature between 1850 and 2100, on a regular 3D grid over the global domain with 192 latitudinal bands, 288 longitudinal bands and 17 pressure levels. The latitude×longitude grid consists of equally spaced data, while the vertical pressure levels are of 1000, 925, 850, 700, 600, 500, 400, 300, 250, 200, 150, 100, 70, 50, 30, 20, and 10 hPa (Taylor et al., 2012), which span from the Earth's surface to the upper stratosphere. The model is solved in sigma coordinates (Pielke, 2002) and then mapped into gridded coordinates, implying that, near the Earth's surface, some temperature values are not defined since over land, especially in mountainous regions, some pressure levels are not present. Therefore, the grid is incomplete for high pressure levels. In Section 4.1,

we discuss how we account for the missing values and other preprocessing aspects of this work. Figure 1 shows an example of realization of the temperature field for 2014.

Since a preliminary analysis has shown that the statistical characteristics of the narrow bands near the poles (especially Antarctica) are very different, we removed the data for the Antarctic continent (South of -62° latitude) and the data North of 82° latitude, at all heights. The total number of latitude bands considered in this analysis is therefore 155, and the data set consists of 1.1×10^9 points.

3 Statistical models for a climate ensemble

In this work, we operate under the assumption that since the runs in the ensemble have different initial conditions, they are statistically independent. This assumption relies on the deterministically chaotic nature of climate models (Lorenz, 1963), although the literature about testing for this assumption is not fully developed (see Collins and Allen (2002); Collins (2002); Branstator and Teng (2010) for some exceptions).

Denote by $\mathbf{T}_r$ the temperature process for realization $r = 1, \ldots, R$, by $\boldsymbol{\mu}$ its mean across realizations and by $\boldsymbol{\varepsilon}_r$ the stochastic component of the statistical model. We assume the following model:

$$\mathbf{T}_r = \boldsymbol{\mu} + \boldsymbol{\varepsilon}_r \qquad \boldsymbol{\varepsilon}_r \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}). \tag{1}$$

If we further denote by h_j the pressure level, by L_m the latitude, by ℓ_n the longitude and by t_k the time, where $j = 1, \ldots, J$, $m = 1, \ldots, M$, $n = 1, \ldots, N$, and $k = 1, \ldots, K$ then we have

$$\begin{aligned} \mathbf{T}_r \;=\; & \{\mathbf{T}_r(h_1, L_1, \ell_1, t_1), \ldots, \mathbf{T}_r(h_1, L_M, \ell_1, t_1), \mathbf{T}_r(h_1, L_1, \ell_2, t_1), \ldots, \\ & \mathbf{T}_r(h_2, L_1, \ell_1, t_1), \ldots, \mathbf{T}_r(h_J, L_M, \ell_N, t_K)\}. \end{aligned}$$

The independence assumption in (1) can be assessed pixelwise by first detrending the data, and then computing the $R \times R$ sample correlation matrix to test if the off-diagonal elements are zero. Similarly, a normality test can be performed.

If the independence assumption across the R realizations is valid, then it is possible to have an estimate of $\mathbf{\Sigma}$ that does not depend on $\boldsymbol{\mu}$ using a restricted loglikelihood. The heuristic behind this approach is that from (1) we know that $\mathbf{T}_r - \mathbf{T}_{r'} \sim \mathcal{N}(\mathbf{0}, 2\mathbf{\Sigma})$, and therefore there is no need to parametrize the mean of the model if the only purpose is to estimate $\mathbf{\Sigma}$. It is possible to derive an explicit restricted likelihood form for $\mathbf{D}_r = \mathbf{T}_r - \frac{1}{R}\sum_{r=1}^{R}\mathbf{T}_r$. This idea was first introduced by Castruccio and Stein (2013) for temperatures at the Earth's surface under a single scenario. Suppose that $\mathbf{\Sigma} = \mathbf{\Sigma}(\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is a vector of unknown covariance parameters. Then

Result 1 *Let* $\mathbf{D} = (\mathbf{D}_1^\top, \ldots, \mathbf{D}_R^\top)^\top$. *The restricted loglikelihood for (1) is*

$$\begin{aligned} l(\boldsymbol{\theta}; \mathbf{D}) &= -\tfrac{KJMN(R-1)}{2} log(2\pi) - \tfrac{1}{2}(R-1) log[\det\{\mathbf{\Sigma}(\boldsymbol{\theta})\}] \\ &\quad -\tfrac{1}{2} KJMN log(R) - \tfrac{1}{2}\sum_{r=1}^{R} \mathbf{D}_r^\top \mathbf{\Sigma}(\boldsymbol{\theta})^{-1} \mathbf{D}_r. \end{aligned} \tag{2}$$

Also, the corresponding estimator for $\boldsymbol{\mu}$ *obtained by generalized least squares is* $\hat{\boldsymbol{\mu}} = \frac{1}{R}\sum_{r=1}^{R}\mathbf{T}_r$.

We do not report the proof since it is a straightforward generalization of that in Castruccio and Stein (2013). In this work, all the four steps of the model we present in the next section estimate the parameters by maximizing (2).

4 The statistical model

In this section, we describe the full model for the 3D spatio-temporal temperature field. The model is spectral in space, thereby automatically generating positive definite matrices, and consists of four distinct stages, each one estimating parameters along a new dimension conditional on the

previous stage, and each step consisting of fewer independent fits to a larger subset of the data. This procedure allows a noticeable degree of flexibility as different statistical features of the data can be estimated independently by multiple processors and merged subsequently. In Section 4.1 we discuss some preprocessing aspects before the introduction of the model, while in Sections 4.2, 4.3, 4.4 and 4.5, we present the different stages of the model.

4.1 Preliminaries: missing values, asymptotic standard deviations and computer specifics

As mentioned in Section 2, remapping from sigma coordinates (Pielke, 2002) to gridded coordinates implies that some temperatures are physically inconsistent for high pressure levels, or equivalently for low altitudes. An extreme example is in the Himalaya regions (see Figure 1) where the pressure cannot be 1000 hPa, thus no physical value can be assigned. In this case, the data are assigned the value of 0, since this would be the expected value of $\mathbf{D}_r$ at each location. However, some regions at low altitudes have too many missing values to deliver meaningful results in a statistical analysis. We therefore assign the value of 0 only to latitude bands with at least 20/280 defined temperatures, otherwise we discard the entire band.

Given the considerably large size of the data set, many of the parameters' asymptotic standard deviations are orders of magnitude smaller than the point estimates. Further, since the statistical model comprises of millions of parameters, we decide not to report the uncertainty of the estimates throughout this paper. Nevertheless, the computational time reported comprises of the Hessian calculation at the optimum and the storage of the asymptotic standard deviations. Without the evaluation of the Fisher information, approximately a day of computation can be saved.

In terms of computational requirements, for this analysis the number of processors and the RAM size are the most important features. We use a workstation with two twelve-cores Intel Xeon

E5-2697 v2 (at nominal frequency 2.7Ghz) and 200 Gb of RAM memory, and all the steps were executed in MATLAB with the NelderMead minimization algorithm.

4.2 Step 1: temporal part

Denote by $\boldsymbol{\varepsilon}(t;r)$ the $JMN \times 1$ vector of the stochastic component for realization r and time t. We assume an autoregressive AR(2) structure with separate parameters for every location:

$$\begin{aligned} \boldsymbol{\varepsilon}(t;r) &= \boldsymbol{\varphi}_1\boldsymbol{\varepsilon}(t-1;r) + \boldsymbol{\varphi}_2\boldsymbol{\varepsilon}(t-2;r) + \boldsymbol{\eta}(t;r), \\ \boldsymbol{\eta}(t;r) &\overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{SCS}), \end{aligned} \tag{3}$$

where $\boldsymbol{\varphi}_1$ and $\boldsymbol{\varphi}_2$ are $JMN \times JMN$ diagonal matrices with the autoregressive coefficients, $\mathbf{S}$ is a diagonal matrix with the standard deviations for each grid point, and $\mathbf{C}$ is the correlation matrix. The estimation of $\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2$ and $\mathbf{S}$ is performed independently across the $\mathbf{D}_1, \ldots, \mathbf{D}_R$ and then averaged.

This procedure requires approximately 7.6×10^5 independent fits of 6 time series of 251 year. It does not require storage of large matrices and can be performed in approximately 4.6 hours on the workstation. By allowing different locations to have different parameters, the model is able to capture different temporal patterns. Equation (3) assumes that an AR(2) structure is sufficient to capture the temporal features of the data (see supplements for diagnostic and discussion). Moreover, (3) assumes no temporal cross-correlation between neighboring points (see the supplements) and stationarity across time, discussed in a similar context in Castruccio and Stein (2013). If the assumption of the temporal structure (3) is not adequate, the parameter estimates are not informative and therefore the statistical model would not be able to compress effectively the data.

In Figure 2, the results of the fit for 925hPa are reported. The white areas correspond to the mountainous regions where temperature is not defined. Values of $\hat{\boldsymbol{\varphi}}_1$ for this height are mostly positive with the noticeable exception of the Indonesian region, while the values of $\hat{\boldsymbol{\varphi}}_2$ show significant

patterns especially in the Equatorial Pacific region (see supplements for a plot of the p-values). $\hat{\mathbf{S}}^2$ shows higher variance at high latitudes, a well-established feature of climate model projections. We also plot the marginal variance $\hat{\mathbf{S}}^2/(\mathbf{I} - \hat{\boldsymbol{\varphi}}_1^2 - \hat{\boldsymbol{\varphi}}_2^2)$ (where the division sign is componentwise), which shows qualitatively similar patterns to $\hat{\mathbf{S}}^2$, but differs significantly in magnitude.

We define $\mathbf{H}(t;r) = \{\mathbf{D}(t;r) - \boldsymbol{\varphi}_1\mathbf{D}(t-1;r) - \boldsymbol{\varphi}_2\mathbf{D}(t-2;r)\}\,\mathbf{S}^{-1}$, where $\mathbf{D}(t;r)$ is the 3D field for realization r and time t. We also define $\hat{\mathbf{H}}(t;r)$ as the same expression as above with the estimated AR(2) parameters. The next sections describe a model for $\mathbf{C}$ that can be estimated by $\hat{\mathbf{H}}(t;r)$.

4.3 Step 2: single band

Conditional on the previous step, we describe a model for the spatial correlation of $\boldsymbol{\eta}(t;r)$ at different longitudes but at the same latitude and height. The points are equally spaced on a circle: assuming stationarity across longitudes results in an exactly circulant covariance matrix (Davis, 1979) and therefore independence across wavenumbers in the spectral domain (see Castruccio and Genton (2014) for a full discussion on the stationarity assumption). It is thus natural to model the spectral density of $\boldsymbol{\eta}_{j,m} = \{\boldsymbol{\eta}(h_j, L_m, \ell_1), \ldots, \boldsymbol{\eta}(h_j, L_m, \ell_N)\}^\top$ for fixed j and m (temporal and realization indices are omitted since from (3) we are assuming the same distribution). Denote by $\tilde{\boldsymbol{\eta}}_{j,m}$ the band-wise Fourier transform; we propose the following model for $\mathrm{var}(\tilde{\boldsymbol{\eta}}_{j,m})$ at wavenumber c:

$$f_{j,m}(c;\boldsymbol{\theta}_{j,m}) = \begin{cases} v_{0;j,m} & \text{if } c = 0, \\ v_{n;j,m} & \text{if } c = n \text{ or } c = N-n+1, n \le V_{j,m}, \\ \frac{\phi_{j,m}}{\left\{\alpha_{j,m}^2 + 4\sin^2\left(\frac{c}{N}\pi\right)\right\}^{\nu_{j,m}+1/2}} & \text{otherwise,} \end{cases} \tag{4}$$

for $\boldsymbol{\theta}_{j,m} = (V_{j,m}, v_{0;j,m}, \ldots, v_{V_{j,m};j,m}, \phi_{j,m}, \nu_{j,m}, \alpha_{j,m})^\top$ and $c = 0, \ldots, N-1$. The choice of the polynomial decay in (4) describes a Matérn-like correlation, modified to allow a smooth transition at high wavenumbers (Castruccio and Stein, 2013).

The latitudes exhibit an increasing smoothness as the height increases. This feature makes the inference of the spatial correlation challenging (Stein, 1999) especially at the last height levels for the equatorial regions. To improve the fit, we allow the first $V_{j,m} + 1$ wavenumbers to have separate values. We choose $V_{j,m} = 6$ for all j if $1 \leq j < 14$ and if $L_m \leq -23^\circ$ or $L_m \geq 23^\circ$, and $j = 15, 16, 17$ while we choose $V_{j,m} = 13$ if $-23^\circ < L_m < 23^\circ$ and $j > 14$. A preliminary study has shown how the $\hat{v}_{0;j,m}, \ldots, \hat{v}_{V_{j,m}}$ are always very close to the empirical periodogram estimates. Therefore, to improve efficiency, we estimate only $(\phi_{j,m}, \nu_{j,m}, \alpha_{j,m})$ via restricted likelihood (1), conditional on estimating the low wavenumbers via an empirical periodogram.

This step requires approximately 3200 independent fits using (2) for a data set of size $288 \times 251 \times 6$ representing the band×time across realizations. Each likelihood evaluation requires around 0.25 seconds, and the workstation performs the fit in approximately 4.4 hours.

Figure 3 shows the results of the fit for some of the heights and latitude bands. Figures 3(a) and 3(b) show the comparison of the fitted periodogram of $\hat{\mathbf{H}}(t; r)$ for two latitudes with the empirical nonparametric estimate, averaged over realizations and times. The low wavenumbers are identical by construction, but model (4) is flexible enough to capture the high wavenumbers as well. In Figures 3(c) and 3(d), the corresponding correlation functions are shown. For both altitudes, the fitted and empirical correlations look almost indistinguishable, mostly because the large-scale feature of the curve is determined by the low wavenumbers, which are the same. In Figures 3(e) and 3(f), we see a comparison of the fitted and empirical squared difference of $\hat{\mathbf{H}}(t; r)$ between two neighboring points at the same latitude and height, averaged over longitude, times, and realization (details can be found in the supplements). Different pressure levels show similar results. We therefore conclude that the model is able to capture this high wavenumber feature of the data. It is noticeable how, for high altitudes, the averaged squared contrast is significantly smaller than for the lower altitudes, as the temperature field at high latitudes is noticeably smoother.

The plots of $\hat{\phi}_{j,m}, \hat{\alpha}_{j,m}$ and $\hat{\nu}_{j,m}$ for $p = 850$hPa ($j = 3$) and $p = 20$hPa ($j = 16$) are reported in the supplement.

4.4 Step 3: multiple latitudes

Once the single band parameters have been estimated, a model for $(\tilde{\boldsymbol{\eta}}_{j,1}, \ldots, \tilde{\boldsymbol{\eta}}_{j,M})^\top$ is defined. We assume that

$$\operatorname{cov}\{\tilde{\boldsymbol{\eta}}_{j,m}(c), \tilde{\boldsymbol{\eta}}_{j,m'}(c')\} = 0, \qquad \text{for all } m \neq m', c \neq c'. \tag{5}$$

This allows for some degree of sparsity in the spectral domain as only the coherence across processes sharing the same wavenumber needs to be defined. When $c = c'$ in (5) we assume that

$$\begin{aligned} \left|\operatorname{corr}(\tilde{\boldsymbol{\eta}}_{j,m}, \tilde{\boldsymbol{\eta}}_{j,m'})(c; \xi_j, \tau_j)\right| &= \operatorname{Re}\left\{\hat{\mathbf{V}}_j(c)\right\} I(c \leq 7) + \left[\frac{\xi_j}{\{1+4\sin^2(\frac{c}{N}\pi)\}^{\nu_j+1/2}}\right]^k I(c > 7), \\ \arg\left\{\operatorname{corr}(\tilde{\boldsymbol{\eta}}_{j,m}, \tilde{\boldsymbol{\eta}}_{j,m'})(c)\right\} &= 0, \end{aligned} \tag{6}$$

where k is the number of bands separating m and m', Re{} is the real part and $\hat{\mathbf{V}}_j(c)$ is the $M \times M$ empirical coherence at wavenumber c and height j of the process averaged over time and realizations. As in the previous step, this choice of nonparametric estimation stems from the fact that the process is very regular and low wavenumbers are difficult to estimate. The first equation assumes an exponential decay of the coherence across latitude modulated by ξ_j and an exponential decay across wavenumbers modulated by ν_j. The second equation has been shown to be reasonable for data at this time scale (Castruccio and Stein, 2013).

Distributed computing can be used in several ways in this step. The fit for the 17 heights can be done independently, but it is more efficient to distribute the computation of a likelihood for a single height and fit the heights sequentially. We call $\boldsymbol{\Sigma}_c$ the $M \times M$ coherence matrix for wavenumbers c, $\tilde{\hat{\mathbf{H}}}(t; r, c)$ the band-wise Fourier transform of $\hat{\mathbf{H}}(t; r)$ evaluated at $c = 0, \ldots, N-1$, and C a generic

constant. (1) can be written as

$$\begin{aligned} l(\boldsymbol{\theta};\mathbf{D}) &= C - T(R-1)\sum_{c=1}^{\lfloor N/2\rfloor-1}\log\{\det(\boldsymbol{\Sigma}_c)\} - \tfrac{1}{2}T(R-1)\log\{\det(\boldsymbol{\Sigma}_0)\} \\ &\quad -\tfrac{1}{2}T(R-1)\log\{\det(\boldsymbol{\Sigma}_{\lfloor N/2\rfloor})\} - \sum_{r=1}^{R}\sum_{t=1}^{K}\sum_{c=0}^{N-1}\tilde{\hat{\mathbf{H}}}^{\top}(t;r,c)\boldsymbol{\Sigma}_c^{-1}\tilde{\hat{\mathbf{H}}}(t;r,c). \end{aligned} \tag{7}$$

This allows to compute the logdeterminant by distributing the computations across c for one realization, and then the quadratic form for every r by distributing across t. Every likelihood evaluation requires approximately 12 seconds and the fit for all heights requires 16.7 hours.

In Figure 4, we see the results of the fit for the same two pressure levels as in Figure 3. Panels (a) and (b) show the fit of the cross-periodogram of $\hat{\mathbf{H}}(t;r)$ for neighboring bands at the same height. As in the previous step, the decay is sharper at higher altitudes but the model is able to capture these different behaviors. Panels (c) and (d) show the corresponding cross-correlation and, as in Section 4, the two lines are indistinguishable since the overall shape is mostly determined by the low wavenumbers, which are the same. Panels (e) and (f) show the north-south contrasts (see details in the supplements) and, as before, the values at higher altitudes are smaller since the process is smoother. Overall, the model is able to capture this local variation, but there is some misfit in some parts of the southern hemisphere for low altitudes and in the equatorial region at high altitudes. In the latter case, this is an artifact of the logscale, as the correlation function is essentially constant across the band, as it is evident from panel (d). The values of $\hat{\xi}_j$ and $\hat{\tau}_j$ are reported in the supplements.

4.5 Step 4: multiple heights

Conditional on the previous steps, a model for describing the coherence across multiple heights is then defined. We assume that

$$\begin{aligned} \left|\mathrm{corr}(\tilde{\boldsymbol{\eta}}_{j,m},\tilde{\boldsymbol{\eta}}_{j',m'})(c;\xi_j,\tau_j)\right| &= \mathrm{Re}\left\{\hat{\mathbf{V}}(c)\right\} I(c\le 7) + \zeta^{k'} I(\{m=m'\}\cap\{c>7\}), \\ \arg\left\{\mathrm{corr}(\hat{\boldsymbol{\eta}}_j,\hat{\boldsymbol{\eta}}_{j'})(c)\right\} &= 0, \end{aligned} \tag{8}$$

where k' represents the number of altitude bands separating the two fields. The model assumes a coherence that is exponentially decaying across heights for the same latitude, but not across wavenumbers, since preliminary analyses have shown that a parameter similar to ν_j in Section 4.4 was not needed. The resulting model in the spatial domain assumes that $\mathrm{cov}\{\boldsymbol{\eta}(h_j, L_m, \ell_n), \boldsymbol{\eta}(h_{j'}, L_{m'}, \ell_{n'})\} = f(h_j, h_{j'}, L_m, L_{m'}, \ell_n - \ell_{n'})$, which is a generalization of the *axially symmetric process* (Jones, 1963; Jun and Stein, 2007, 2008; Castruccio and Stein, 2013) with the further constraint of being longitudinally reversible (Stein, 2007).

The likelihood evaluation at this stage is extremely demanding both in terms of flops and storage space. The algorithm we present is not suitable for computers with less than 100 Gb of RAM and could be modified for diminishing the storage space at the expense of more on-the-fly computations, although this would result in an increase in the already long computational time.

We denote by[1] $\boldsymbol{\Sigma}_c$ the $MJ \times MJ$ coherence matrices for wavenumbers c, by $\tilde{\hat{\mathbf{H}}}(t; r, c)$ the bandwise Fourier transform of $\hat{\mathbf{H}}(t; r)$ evaluated at $c = 0, \dots, N-1$ and by C a generic constant. (1) can be written as

$$
\begin{aligned}
l(\boldsymbol{\theta}; \mathbf{D}) &= C - \tfrac{1}{2}T(R-1)\textstyle\sum_{c=0}^{N-1}\log\{\det(\boldsymbol{\Sigma}_c)\} - \tfrac{1}{2}\sum_{c=0}^{N-1}\sum_{r=1}^{R}\sum_{t=1}^{K}\tilde{\hat{\mathbf{H}}}^{\top}(t; r, c)\boldsymbol{\Sigma}_c^{-1}\tilde{\hat{\mathbf{H}}}(t; r, c) \\
&= C - \tfrac{1}{2}\textstyle\sum_{c=0}^{N-1}\Big[T(R-1)\log\{\det(\boldsymbol{\Sigma}_c)\} - \sum_{r=1}^{R}\sum_{t=1}^{K}\tilde{\hat{\mathbf{H}}}_{t,r;c}^{\top}\boldsymbol{\Sigma}_c^{-1}\tilde{\hat{\mathbf{H}}}_{t,r;c}\Big] \\
&= C - \tfrac{1}{2}T(R-1)\textstyle\sum_{c=0}^{\lfloor N/2\rfloor}\big[2\log\{\det(\boldsymbol{\Sigma}_c)\}\mathrm{I}_{0<c<\lfloor N/2\rfloor} \\
&\quad - \textstyle\sum_{r=1}^{R}\sum_{t=1}^{K}\Big(\tilde{\hat{\mathbf{H}}}^{\top}(t; r, c)\boldsymbol{\Sigma}_c^{-1}\tilde{\hat{\mathbf{H}}}(t; r, c) + \tilde{\hat{\mathbf{H}}}^{\top}(t; r, N-c+1)\boldsymbol{\Sigma}_c^{-1}\tilde{\hat{\mathbf{H}}}(t; r, N-c+1)\Big)\Big] \\
&\quad -T(R-1)\tfrac{1}{2}\log\{\det(\boldsymbol{\Sigma}_0)\} - T(R-1)\tfrac{1}{2}\log\{\det(\boldsymbol{\Sigma}_{\lfloor N/2\rfloor})\},
\end{aligned}
\tag{9}
$$

where the last step follows from the circular symmetry, since $\boldsymbol{\Sigma}_c = \boldsymbol{\Sigma}_{N-c+1}$. Stationarity across longitude allows for noticeable saving in storage space, as $\boldsymbol{\Sigma}_c$ for $c = 0, \dots, N-1$ is a $(155\times17)^2\times145$ array, which requires only $\approx$ 10Gb of RAM. Moreover, since every evaluation for c is independent

[1] although the notation is similar to Section 4.4, the matrix size is different since now the covariance comprises of the height component

on the others, (9) can be distributed across at most 145 processors. A single likelihood evaluation requires approximately 15 minutes and the maximization with Hessian evaluation requires approximately 16.4 hours. It is in principle possible to compute the loglikelihood in the form (7), but in order to do that, independent copies of $\mathbf{\Sigma}_c$ should be made, and this would not fit in the RAM memory in this step.

The algorithm estimates $\hat{\zeta} = 0.999$ and in Figure 5 we see the fit for the same altitudes as in Figures 3 and 4. Panels (a) and (b) show the cross-periodogram for two bands at the same latitude and neighboring altitudes. The fit is good and the empirical and fitted periodograms are almost indistinguishable. The corresponding cross-correlations plotted in panels (c) and (d) are also visually indistinguishable because of the nonparametric estimation of $\mathrm{Re}\left\{\hat{\mathbf{V}}(c)\right\}$. The contrasts in panels (e) and (f) (details in the supplements) show some misfit in the northern hemisphere for low altitudes, but for high altitudes the two curves are almost indistinguishable.

In Table 1, a summary of the algorithm is presented. The first step requires six time series to be subsampled, so storage is negligible, but requires millions of fits. Subsequent steps of the algorithm require an increasingly large subsample, thus increasing the storage demand, but with fewer independent fits. The large number of parameters in the last two steps is due to the nonparametric estimation of $\mathrm{Re}\left\{\hat{\mathbf{V}}_j(c)\right\}$ and $\mathrm{Re}\left\{\hat{\mathbf{V}}(c)\right\}$ in (6) and (8). In total, this model requires approximately 27 million parameters, a negligible number compared to the data size ($\approx 2\%$) and less than two days on the dedicated workstation.

5 Simulating the initial condition ensemble

Once the space-time structure has been estimated, a model to reproduce the mean must be defined. In Section 3, we showed how the best estimate for the mean is the average across realizations, so

we compute it and, since the climate is expected to be slowly varying, we fit a smoothing spline to every time series with a mild penalty term of 0.01. We then simulate the model described in Section 4.3, add the fitted spline and compare it with the climate model data; the results are shown in Figure 6.

In Figure 6(a) we see the comparison of six realizations of the climate model output and the statistical model at the nearest point to Jeddah, Saudi Arabia, at approximately 3km above ground level. The two groups of time series exhibit very similar behavior, thus demonstrating how the statistical model can act as an efficient surrogate of the climate model. In the years between 1850 and 2006 the climate model shows sharp drops in three cases, corresponding to the years following major volcanic events in the equatorial area (Krakatoa, 1883; Santa María, 1902 and Mount Pinatubo, 1991). This feature is not captured in the statistical model, although the spline could be adjusted to account for the sharp drop in the year of interest. In Figure 6(b), we see the histogram for the annual temperature of 2014 in the case of the model (left) and for 105 realizations (extracted at the specified location and time from full simulations) of the statistical model. In the case of the model, only six realizations are available and a histogram that could give an idea of the uncertainty would not be possible with current computational facilities. The statistical model instead is able to generate a much more informative histogram with orders of magnitudes more realizations in less than 4 hours. Figure 6(c) shows the comparison between latitudinal bands in the same setting as in Figure 6(a). The overall pattern is captured by the statistical model, as it is difficult to distinguish the gray and red curves. This is even more evident in Figure 6(d), where the vertical profile of temperature is shown for the year 2014 for the grid point near Jeddah. By allowing correlation across altitudes, the statistical model is able to capture the main features of the vertical profile: a drop in the troposphere, the temperature inversion in the tropopause and a further drop in the lower stratosphere.

6 Discussion

In this work, we have shown how a statistical model can be used as an efficient tool for reproducing global annual 3D temperature fields for an initial condition ensemble and therefore compress the data size. Many extensions can be considered. Finer temporal scales such as monthly aggregated data could be modeled by allowing fixed seasonal effects, or by allowing a cyclostationary process (Gardner et al., 2006) as random effect. Besides the foreseen increase in model complexity and computational time due to the twelve-fold increase of the data size, the cross-correlation among sites could be non-negligible, thus requiring a very different modeling strategy. Multiple scenarios could be also accounted for: it has been shown in previous work (Castruccio and Stein, 2013) that different forcings have a similar space-time structure for surface temperature, and we speculate a similar outcome for 3D temperatures, although the diagnostic would be challenging due to the high dimensionality of the parameter space. It is also possible to extend the proposed methodology to cross-correlation for multiple physical variables, or even multiple climate models. Visualization of such a complex data structure has recently been explored using a virtual reality environment and visuanimation in Genton et al. (2015).

Further, different statistical models can also be developed depending on the scientific questions that needs to be addressed by climate scientists. If the goal is to understand the correlation of longitudinal profiles of temperatures, a model that assumes longitudinal stationarity such as the one we propose is likely not optimal. The validation procedure proposed here focuses on the spatio-temporal structure of the data, and other more data-specific criteria can (and should) be proposed, such as reproduction of the patterns in the El Niño Southern Oscillation. However, the state-of-the-art in validating compressed climate data (Baker et al., 2014) has been so far limited to aggregating pointwise discrepancy measures over the spatial domain, and we propose a validation

criteria that is more suitable for spatio-temporally resolved data.

Any statistical model proposed for a large amount of data needs to be always carefully designed to make use of distributed computing and will necessarily entail some degrees of approximation. In this work, the key assumption is that the estimated parameters in the conditional multi-step algorithm are close to the ones obtained with a full likelihood, and that their estimation uncertainty does not propagate significantly throughout the stages. There is partial evidence that the global and conditional optima do not significantly differ in a similar model in Castruccio and Stein (2013), but a similar investigation in this work would require optimizing all 27 million parameters simultaneously, which is not feasible.

This work also shows how powerful computers can be used as an efficient tool by statisticians to provide results of interest to the climate community. Our direction of investigation is not focused on reducing the space/time information to a feasible problem on a laptop, but rather to use more sophisticated hardware and to explore how the "big data" challenge in statistics can be tackled from a different perspective. This perspective has the advantage of not requiring any ad hoc methodologies such as the choice of the basis in a low rank approximation or the choice of blocks in a composite likelihood approach, but it presents different and equally interesting challenges, such as modulating the fitting procedure to be parallelizable (the multi-stage conditional approach we proposed is ideal for gridded data that are stationary in time, but similar strategies can be devised for more complex geometries) and formulating the likelihood allowing efficient matrix storage, a procedure clearly dependent on the computational resources available.

Since the ultimate goal of this work is not inference on the real climate, but rather an efficient reproduction of some features of the climate model, our statistical model makes no attempt to characterize and understand the temperature process and compare it to observational data. Our statistical model reproduces features of the state of the climate only to the extent that the original

climate model does, and we believe a discussion about "structural error" of the climate model from the real state of the climate (Rougier, 2007) is beyond the scope of this work.

We think that the use of distributed computing for space/time analysis is a promising and unexplored direction and more effort should be devoted to understanding how current computational facilities can change our approach to big data problems and to serve areas of science where the amount of data has exponentially increased in the last decade. Our future research will include adapting scalable algorithms to satellite retrievals but also the use of hybrid codes with different languages at different stages of the analysis, and the use of Graphical Processing Units to further distribute the computational load.

Acknowledgements

Research reported in this publication was supported by the King Abdullah University of Science and Technology (KAUST). The authors would like to thank the KAUST Supercomputing Laboratory and the KAUST Information Technology group for having provided the computational resources and the technical support needed for this work. We acknowledge the World Climate Research Programme's Working Group on Coupled Modeling, which is responsible for CMIP, and we thank NCAR for producing and making available their model output. For CMIP, the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provided coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals.

References

Anitescu, M., Chen, J., and Wang, L. (2012), "A Matrix-free Approach for Solving the Gaussian Process Maximum Likelihood Problem," *SIAM Journal of Scientific Computing*, 34.

Baker, A., Xu, H., Dennis, J., 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*, New York, NY, USA: ACM, HPDC '14, pp. 203–214.

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008), "Gaussian Predictive Process Models for Large Spatial Data Sets," *Journal of the Royal Statistical Society: Series B*, 70, 825–848.

Bhat, K., Haran, M., Olson, R., and Keller, K. (2012), "Inferring Likelihoods and Climate System Characteristics from Climate Models and Multiple Tracers," *Environmetrics*, 23, 345–362.

Bicer, T., Jian, Y., Chiu, D., Agrawal, G., and Schuchardt, K. (2013), "Integrating Online Compression to Accelerate Large-Scale Data Analytics Applications," in *Parallel Distributed Processing (IPDPS), 2013 IEEE 27th International Symposium on*, pp. 1205–1216.

Branstator, G. and Teng, H. (2010), "Two Limits of Initial-value Decadal Predictability in a CGCM," *Journal of Climate*, 23, 6292–6311.

Burtscher, M. and Ratanaworabhan, P. (2007), "High Throughput Compression of Double-Precision Floating-Point Data," in *Data Compression Conference, 2007. DCC '07*, pp. 293–302.

Castruccio, S. and Genton, M. G. (2014), "Beyond Axial Symmetry: An Improved Class of Models for Global Data," *Stat*, 3, 48–55.

Castruccio, S., McInerney, D. J., Stein, M. L., Liu, F., Jacob, R. J., and Moyer, E. J. (2014), "Statistical Emulation of Climate Model Projections Based on Precomputed GCM Runs," *Journal of Climate*, 27, 1829–1844.

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

Chang, W., Haran, M., Olson, R., and Keller, K. (2015), "A Composite Likelihood Approach to Computer Model Calibration using High-dimensional Spatial Data," *Statistica Sinica*, 25, 243–260.

Collins, M. (2002), "Climate Predictability on Interannual to Decadal Time Scales: the Initial Value Problem," *Climate Dynamics*, 19, 671–692.

Collins, M. and Allen, M. R. (2002), "Assessing the Relative Roles of Initial and Boundary Conditions in Interannual to Decadal Climate Predictability," *Journal of Climate*, 15, 3104–3109.

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

Davis, P. (1979), *Circulant Matrices*, New York: Wiley.

Drignei, D., Forest, C. E., and Nychka, D. (2008), "Parameter Estimation for Computationally Intensive Nonlinear Regression with an Application to Climate Modeling," *Annals of Applied Statistics*, 2, 1217–1230.

Eidsvik, J., Shaby, B. A., Reich, B. J., Wheeler, M., and Niemi, J. (2014), "Estimation and Prediction in Spatial Models With Block Composite Likelihoods," *Journal of Computational and Graphical Statistics*, 23, 295–315.

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

Gardner, W., Napolitano, A., and Paura, L. (2006), "Cyclostationarity: Half a Century of Research," *Signal Processing*, 86, 639 – 697.

Gent, P. R. et al. (2011), "The Community Climate System Model version 4," *Journal of Climate*, 24, 4973–4991.

Genton, M. G., Castruccio, S., Crippa, P., Dutta, S., Huser, R., Sun, Y., and Vettori, S. (2015), "Visuanimation in statistics," *Stat*, in press.

Gomez, L. and Cappello, F. (2013), "Improving Floating Point Compression through Binary Masks," in *IEEE BigData 2013*, Santa Barbara, California.

Higdon, D. (1998), "A Process-convolution Approach to Modelling Temperature in the North Atlantic Ocean," *Environmental and Ecological Statistics*, 5, 173–190.

Holden, P. B. and Edwards, N. R. (2010), "Dimensionally Reduced Emulation of an AOGCM for Application to Integrated Assessment Modelling," *Geophysical Research Letters*, 37.

Holden, P. B., Edwards, N. R., Garthwaite, P. H., Fraedrich, K., Lunkeit, F., Kirk, E., Labriet, M., Kanudia, A., and Babonneau, F. (2013), "PLASIM-ENTSem: a Spatio-temporal Emulator of Future Climate Change for Impacts Assessment," *Geoscientific Model Development Discussions*, 6, 3349–3380.

Hübbe, N., Wegener, A., Kunkel, J., Ling, Y., and Ludwig, T. (2013), "Evaluating Lossy Compression on Climate Data," in *Supercomputing*, eds. Kunkel, J., Ludwig, T., and Meuer, H., Springer Berlin Heidelberg, vol. 7905 of *Lecture Notes in Computer Science*, pp. 343–356.

Jones, R. (1963), "Stochastic Processes on a Sphere," *The Annals of Mathematical Statitistics*, 34, 213–218.

Jun, M. and Stein, M. (2007), "An Approach to Producing Space-time Covariance Functions on Spheres," *Technometrics*, 49, 468–479.

— (2008), "Nonstationary Covariance Models for Global Data," *Annals of Applied Statistics*, 2, 1271–1289.

Lakshminarasimhan, S., Shah, N., Ethier, S., Klasky, S., Latham, R., Ross, R., and Samatova, N. (2011), "Compressing the Incompressible with ISABELA: In-situ Reduction of Spatio-temporal Data," in *Proceedings of the 17th International Conference on Parallel Processing - Volume Part I*, Berlin, Heidelberg: Springer-Verlag, Euro-Par'11, pp. 366–379.

Laney, D., Langer, S., Weber, C., Lindstrom, P., and Wegener, A. (2013), "Assessing the Effects of Data Compression in Simulations Using Physically Motivated Metrics," in *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, New York, NY, USA: ACM, SC '13, pp. 76:1–76:12.

Lindgren, F., Rue, H., and Lindström, J. (2011), "An Explicit Link Between Gaussian fields and Gaussian Markov Random Fields: the Stochastic Partial Differential Equation Approach," *Journal of the Royal Statistical Society: Series B*, 73, 423–498.

Lindstrom, P. and Isenburg, M. (2006), "Fast and Efficient Compression of Floating-Point Data," *Visualization and Computer Graphics, IEEE Transactions on*, 12, 1245–1250.

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

Meehl, G. A., Stocker, T., Collins, W., Friedlingstein, P., Gaye, A., Gregory, J. M., Kitoh, A., Knutti, R., Murphy, J., Noda, A., Raper, S., Watterson, I., Weaver, A., and Zhao, Z.-C. (2007), *Global Climate Projections. In climate change 2007: The Physical Sciences Basis. Contribution of Working Group I to the Forth Assessment report of the Intergovernmental Panel on Climate Change*, Cambridge University press, Cambridge, United Kingdom and New York, NY, USA: S. Solomon, Qin, D. and Manning, M. and Chen, Z. and Marquis, M. and Averyt, K.B. and Tignor, M. and Miller, H.L. (eds).

Pielke, R. (2002), *Mesoscale Meteorological Modeling*, International Geophysical Series, Volume 38.

Rougier, J. (2007), "Probabilistic Inference for Future Climate Using an Ensemble of Climate Model Evaluations," *Climatic Change*, 81, 247–264.

Rue, H. and Held, L. (2005), *Gaussian Markov Random Fields Theory and Applications*, Chapman & Hall/CRC.

Sansó, B. and Forest, C. (2009), "Statistical Calibration of Climate System Properties," *Journal of the Royal Statistical Society: Series C*, 58, 485–503.

Sansó, B., Forest, C., and Zantedeschi, D. (2008), "Inferring Climate System Properties Using a Computer Model," *Bayesian Analysis*, 3, 1–37.

Schendel, E., Jin, Y., Shah, N., Chen, J., Chang, C., Ku, S.-H., Ethier, S., Klasky, S., Latham, R., Ross, R., and Samatova, N. (2012), "ISOBAR Preconditioner for Effective and High-throughput Lossless Data Compression," in *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*, pp. 138–149.

Simpson, D., Lindgren, F., and Rue, H. (2012), "In Order to Make Spatial Statistics Computationally Feasible, we Need to Forget about the Covariance Function," *Environmetrics*, 23, 65–74.

Stein, M. (1999), *Interpolation of Spatial Data: Some Theory for Kriging*, New York: Springer.

— (2007), "Spatial Variation of Total Column Ozone on a Global Scale," *Annals of Applied Statistics*, 1, 191–210.

Stein, M., Chen, J., and Anitescu, M. (2012), "Difference Filter Preconditioning for Large Covariance Matrices," *SIAM Journal on Matrix Analysis and Applications*, 33.

Stein, M., Chi, Z., and Welty, L. J. (2004), "Approximating likelihoods for large spatial data sets," *Journal of the Royal Statistical Society: Series B*, 66, 275–296.

Stein, M. L. (2014), "Limitations on Low Rank Approximations for Covariance Matrices of Spatial Data," *Spatial Statistics*, 8, 1–19.

Sun, Y., Li, B., and Genton, M. G. (2012), *Geostatistics for Large Datasets*, Porcu, E., Montero, J. M., Schlather, M., Springer, chap. 3, pp. 55–77.

Taylor, K., Stouffer, R., and Meehl, G. (2012), "An Overview of CMIP5 and the Experiment Design," *Bulletin of the American Meteorological Society*, 93, 485–498.

Van Vuuren, D. et al. (2011), "The Representative Concentration Pathways: an Overview," *Climatic Change*, 109, 5–31.

Vecchia, A. V. (1988), "Estimation and Model Identification for Continuous Spatial Processes," *Journal of the Royal Statistical Society: Series B*, 50, 297–0312.

Woodring, J., Mniszewski, S., Brislawn, C., DeMarle, D., and Ahrens, J. (2011), "Revisiting wavelet compression for large-scale climate data using JPEG 2000 and ensuring data precision," in *Large Data Analysis and Visualization (LDAV), 2011 IEEE Symposium on*, pp. 31–38.

Xu, G., Liang, F., and Genton, M. (2015), "A Bayesian Spatio-temporal Geostatistical Model with an Auxiliary Lattice for Large Datasets," *Statistica Sinica*, 25, 61–79.

Table 1: Synopsis of the algorithm steps.

step	#fits	subsample size	#parameters	comput. time
time	7.6×10^5	1.5×10^3	2.2×10^6	$\approx$ 4.6h
longitude	2.6×10^3	1.2×10^6	2.7×10^4	$\approx$ 4.4h
latitude	17	6.7×10^7	1.4×10^6	$\approx$ 16.7h
altitude	1	1.1×10^9	2.4×10^7	$\approx$ 16.4h
total	7.6×10^5	1.1×10^9	2.7×10^7	$\approx$ 42.2h

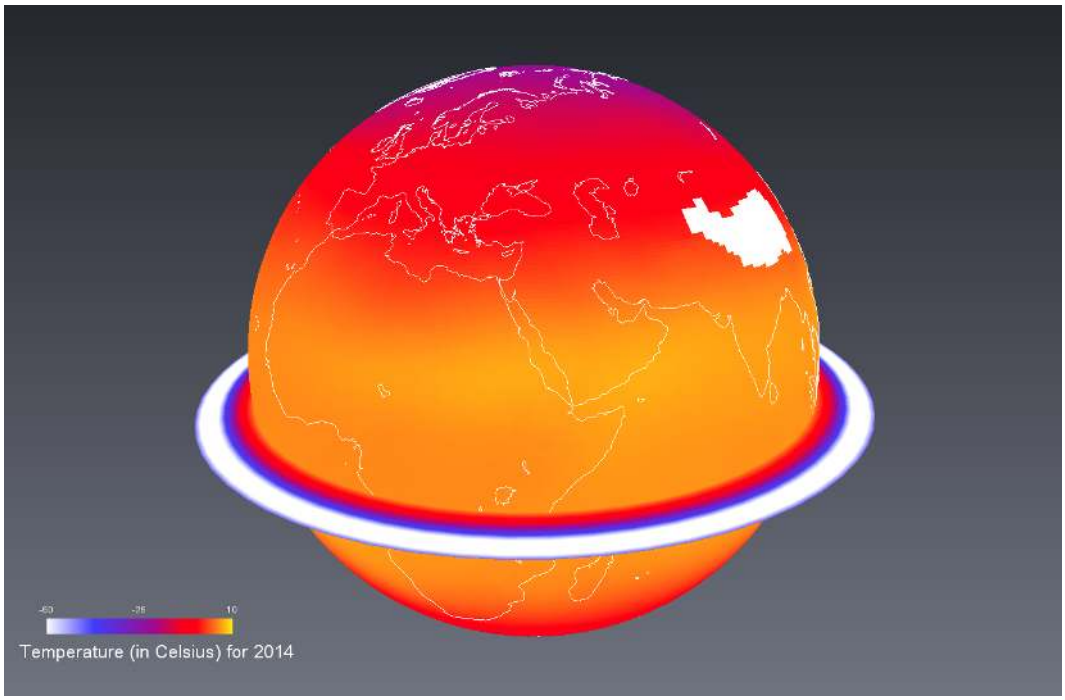

Figure 1: An example of one realization in the ensemble. The temperature field is plotted for 2014 at 925hPa and the vertical profile of temperature for a latitude band is also shown (pressure levels are on a log scale to magnify the effects on the troposphere). The white areas represent points where temperatures are not defined for that pressure level (in this case, the Himalaya region). Some of the main features of 3D global temperature fields are evident, such as colder temperatures for higher latitudes and temperature inversion in the tropopause.

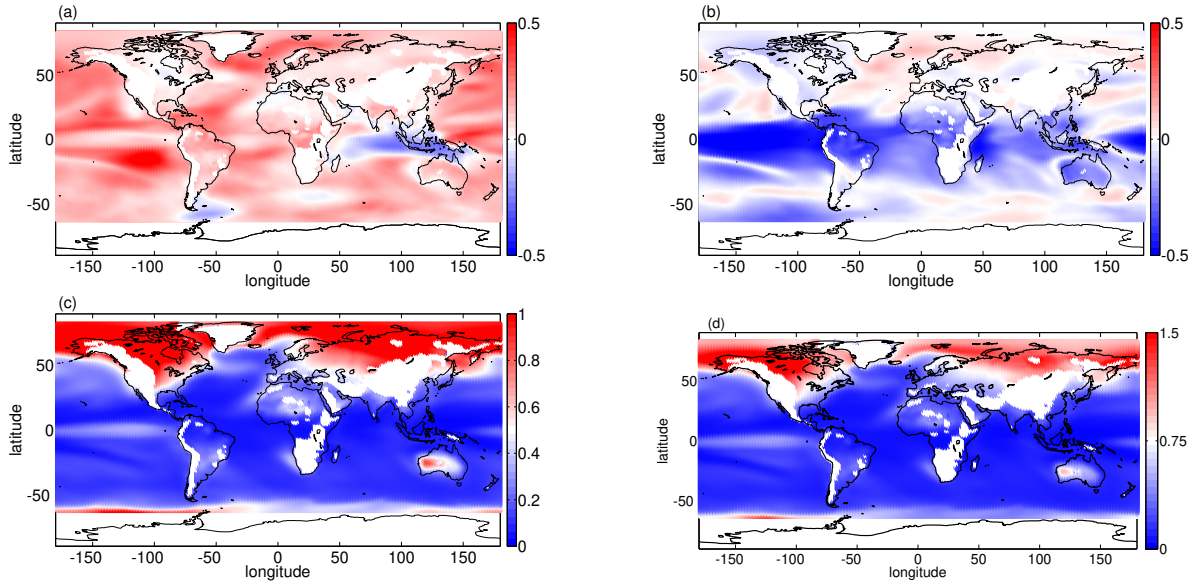

Figure 2: Plots of the estimated autoregressive parameters in (3) for $j = 2$, the second height level. (a): $\hat{\boldsymbol{\varphi}}_1$, (b): $\hat{\boldsymbol{\varphi}}_2$, (c): $\hat{\mathbf{S}}^2$ and (d): the marginal variance $\hat{\mathbf{S}}^2/(\mathbf{I} - \hat{\boldsymbol{\varphi}}_1^2 - \hat{\boldsymbol{\varphi}}_2^2)$.

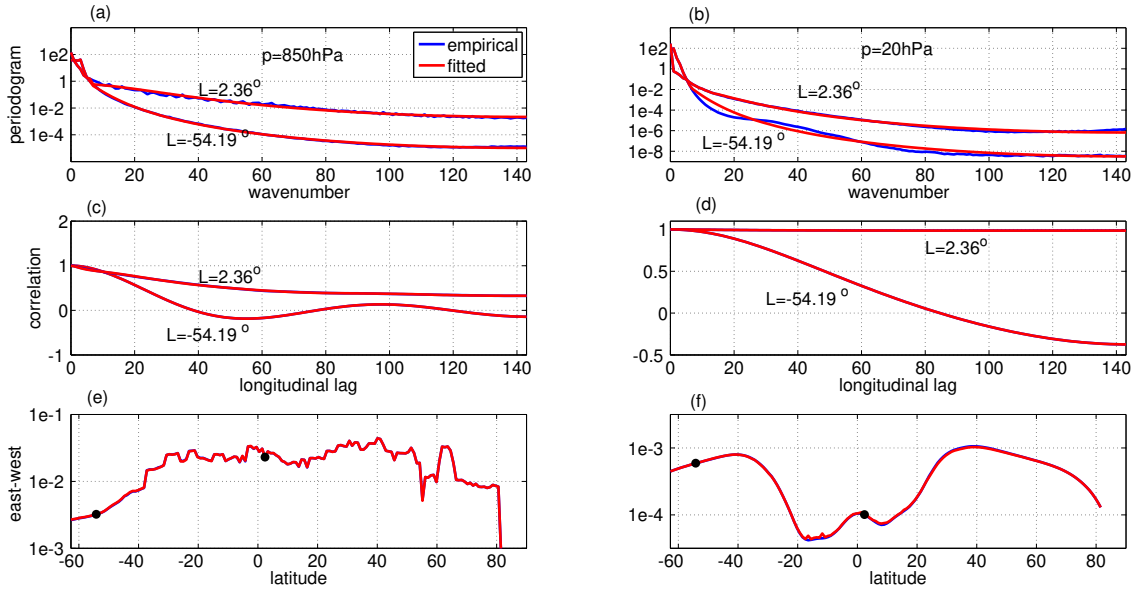

Figure 3: Fit and diagnostic for the single band parameters for p = 850hPa (a,c,e) and p = 20hPa (b,d,f). Panels (a) and (b) show the fit of the periodogram for two bands, panels (c) and (d) the corresponding empirical and fitted correlation function and panels (e) and (f) the empirical and fitted east-west contrast, averaged across longitudes, times and realizations. The black dots correspond to the latitude bands chosen for the results in the above panels.

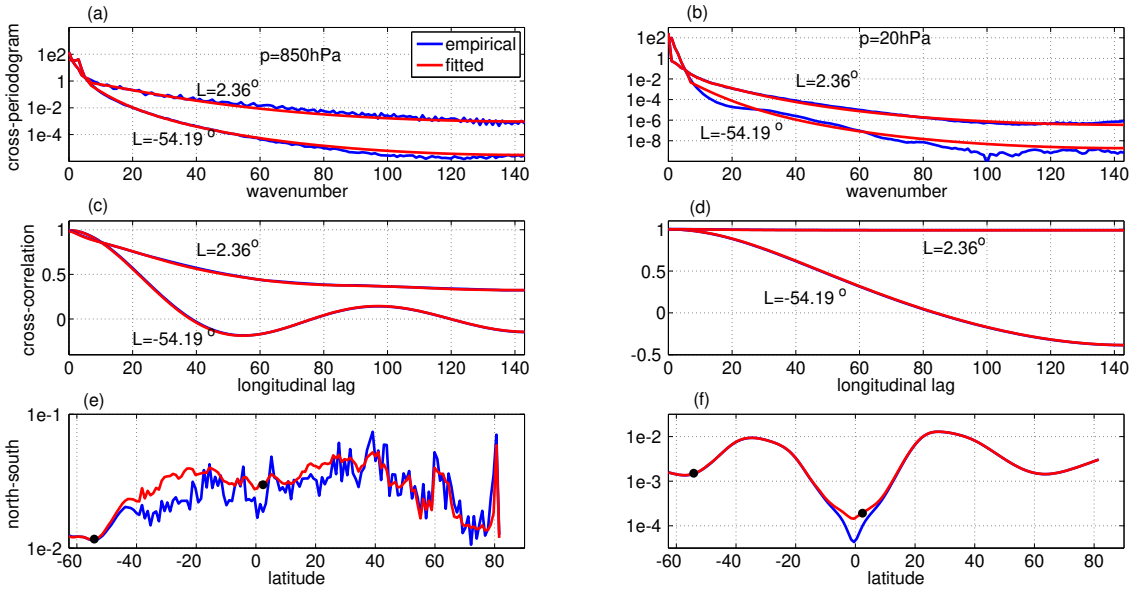

Figure 4: Fit and diagnostic for the multiple band parameters for $p = 850$hPa (a,c,e) and $p = 20$hPa (b,d,f). Panels (a) and (b) show the fit of the cross-periodogram for two bands at same height and longitude and neighboring latitudes, panels (c) and (d) the corresponding empirical and fitted cross-correlation and panels (e) and (f) the empirical and fitted up-down contrast, averaged across longitudes, times and realizations. The black dots correspond to the latitude bands chosen for the results in the above panels.

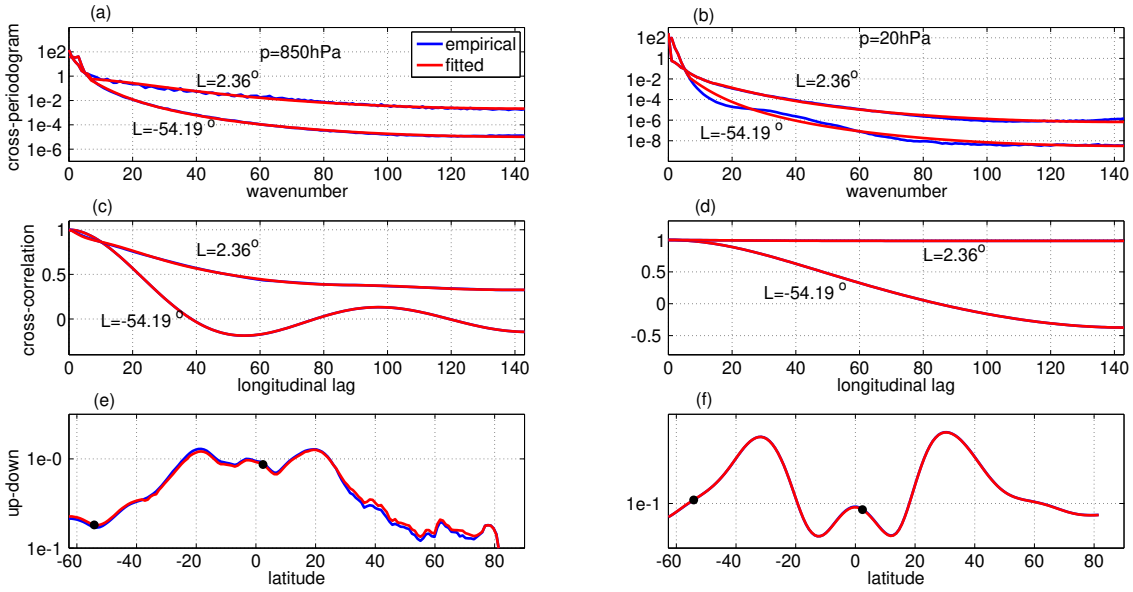

Figure 5: Fit and diagnostic for the multiple band parameters for $p = 850$hPa (a,c,e) and $p = 20$hPa (b,d,f). Panels (a) and (b) show the fit of the cross-periodogram for two bands at same latitude and longitude and neighboring heights, panels (c) and (d) the corresponding empirical and fitted cross-correlation and panels (e) and (f) the empirical and fitted up-down contrast, averaged across longitudes, times and realizations. The black dots correspond to the latitude bands chosen for the results in the above panels.

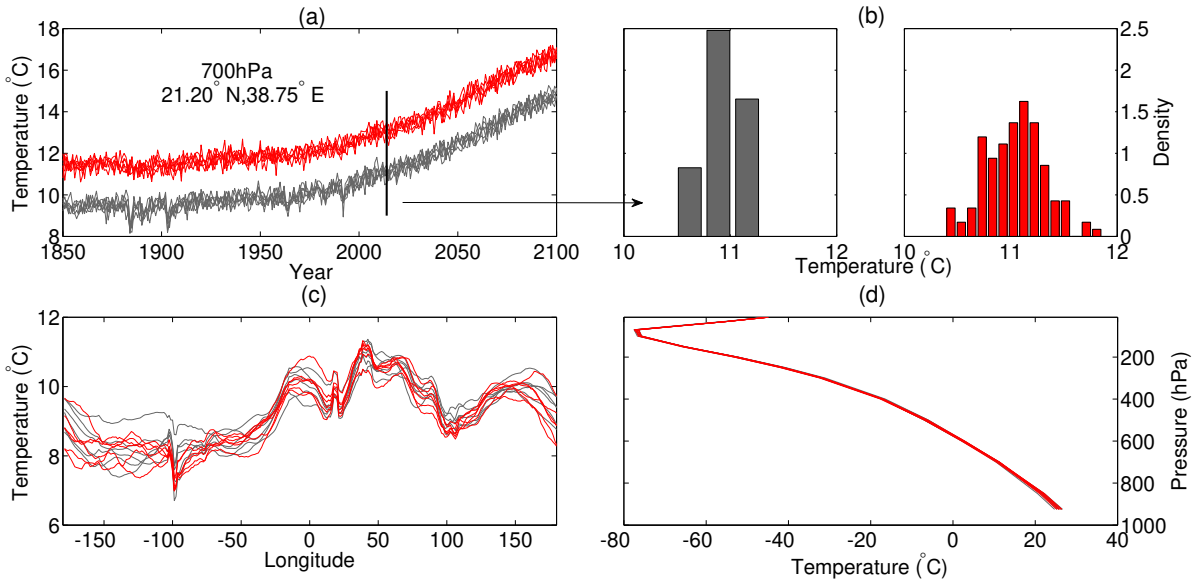

Figure 6: Comparison of climate model output and simulated output. (a) For the nearest point to Jeddah, Saudi Arabia, at approximately 3km above ground level, the $R = 6$ realizations of the model output (in gray) and 6 realizations of the simulated output (in red, offset by 2° C) are shown. (b) For the same setting as (a), we focus on the year 2014 and show a histogram of the values of annual temperature for the climate model (6 realizations) and the statistical model (extracted from 105 full space-time resolved realizations). (c) Comparison of latitudinal bands, with the same setting, latitude and height as in (a) for the year 2014. (d) Comparison of vertical profiles, with the same setting, latitude and longitude as in (a) and for the year 2014.