Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2018-60
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 29 March 2018

Geoscientific Model Development Discussions
Open Access

Computationally Efficient Emulators for Earth System Models

Robert Link[1], Cary Lynch[1], Abigail Snyder[1], Corinne Hartin[1], Ben Kravitz[1], and Ben Bond-Lamberty[1]

[1]Pacific Northwest National Laboratory, Joint Global Change Research Institute, 5825 University Research Ct., College Park, MD

Correspondence: Robert Link (robert.link@pnnl.gov)

Abstract. Earth System Models (ESMs) are the gold standard for producing future projections of climate change, but running them is difficult and costly, and thus researchers are generally limited to a small selection of scenarios. This paper presents a technique for detailed emulation of Earth System Model (ESM) temperature output, based on constructing a deterministic model for the mean response to global temperature. The residuals between the mean response and the observed temperature fields are used to construct variability fields that are added to the mean response to produce the final product. The method produces grid-level output with spatially and temporally coherent variability. Output fields include random components, so the system may be run as many times as necessary to produce large ensembles of fields for uncertainty studies and similar uses. We describe the method, show example outputs, and present statistical verification that it reproduces the ESM properties it is intended to capture. This method, available as an open-source R package, should have utility in the study of climate uncertainty and variability, extreme events, and climate change mitigation.

1 Introduction

There are a variety of scientific applications that need many realizations of one or more future climate scenarios. One prominent example is uncertainty studies, in which the multiple realizations are used to compute a statistical distribution of outcomes (Murphy et al., 2004; Falloon et al., 2014; Sanderson et al., 2015; Bodman and Jones, 2016; Rasmussen et al., 2016). Studying the effects of extreme events, which by definition occur infrequently in any single scenario run, is another example (Greenough et al., 2001), and the study of climate variability is a third (Kay et al., 2015).

Earth System Models (ESMs) are the gold standard for producing future projections of climate change. However, running ESMs is difficult and costly. As a result, most researchers are forced to rely on public libraries of ESM runs produced in model intercomparison projects, such as the CMIP5 (Coupled Model Intercomparison Project) archive (Taylor et al., 2012). Although a few experiments have tried to produce larger ensembles of runs (e.g. Kay et al., 2015), typically users are limited to a small selection of scenarios with only a handful of runs for each scenario. This limited selection may be inadequate for many types of studies. Users might need customized scenarios following some specific future climate pathway not covered by the scenario library, or the small collection of runs might be insufficient for a robust statistical analysis. In these situations, researchers typically turn to *emulators* to get access to a sufficient quantity of data without having to do an infeasible amount of computation.

Climate model emulators attempt to approximate the output a climate model *would have* produced had it been run for a specified scenario. Perhaps the best known emulator algorithm is *pattern scaling*, which develops in each grid cell a linear relationship between global mean temperature T_g and the climate variable or variables being modeled (Mitchell et al., 1999; Mitchell, 2003; Tebaldi and Arblaster, 2014). A variety of enhancements to this basic procedure have been proposed, mostly centering around adding additional predictor variables (*i.e.*, besides just T_g) (MacMartin and Kravitz, 2016), adding nonlinear terms to the emulator function (Neelin et al., 2010), or separating the climate state into components, each with its own dependence on the predictor variables (Holden and Edwards, 2010).

Most of these methods are deterministic functions of their inputs, and thus their outputs can be viewed as expectation values for the ESM output. Real ESM output, however, would have some distribution around these mean response values. We will refer to these departures from the mean response generically as "variability." There have been some attempts to add variability to emulators, but producing realistic variability is difficult, due to the complicated correlation structure exhibited by climate model output over both space and time. Typically methods deal with this difficulty by either placing *a priori* limits on the form of the correlation function (Castruccio and Stein, 2013), or by using bootstrap resampling of existing ESM output (Osborn et al., 2015; Alexeeff et al., 2016).

In this paper we describe a computationally-efficient method for producing climate scenario realizations with realistic variability. The realizations are constructed so as to have the same variance and time-space correlation structure as the ESM data used to train the system. The variability produced by the method includes random components, so the system may be run many times with different random number seeds to produce an ensemble of independent realizations. The results in this study are limited to temperature output at annual resolution. Future papers will extend the method to additional output variables, such as precipitation, and to subannual time resolution.

2 Method

2.1 Notation

In the text that follows, we use uppercase bold symbols (*e.g.* $\mathbf{R}$) to refer to matrices. Vectors are represented using Dirac notation, $|x\rangle$ for a column vector and $\langle x|$ for a row vector. Thus, we can represent a matrix-vector product as $\mathbf{R}|x\rangle$ and a dot product of two vectors as $\langle x\,|y\rangle$.

Occasionally we will add a matrix and a vector; *e.g.*, $\mathbf{B} = \mathbf{A} + |x\rangle$. This should be interpreted to mean that the vector $|x\rangle$ is to be added to each row of the matrix $\mathbf{A}$. Therefore, the length of $|x\rangle$ must be equal to the number of columns in $\mathbf{A}$. This *broadcast* convention is slightly nonstandard mathematically, but it is common in programming languages that support matrix arithmetic (*e.g.* the *numpy* package for python), and simplifies certain expressions that will come up in the derivation.

2.2 Input

Our method requires a collection of ESM model output to train on. Any model can be used, and by switching out the input data the method can be tuned to produce results representative of any desired ESM. For all of the results in this paper we have used the CESM(CAM5) (Community Earth System Model (Community Atmosphere Model)) output from the CMIP5 archive (Taylor et al., 2012). We used surface temperature data from all available 21st century runs for all four Representative Concentration Pathway (RCP) emissions scenarios (RCP2.6, RCP4.5, RCP6.0, and RCP8.5), for a total of 9 runs, each 95 years in length. These data were averaged to annual resolution, for a total of 855 observations of the global temperature state. (We refer to the ESM data as “observations” because they are the data we are trying to emulate. Similarly, when we refer to “results” or “model output” without further clarification, we are talking about the data produced by the emulator.)

Throughout the discussion, we will treat each temperature state observation as a vector, with each grid cell providing one entry in the vector. The ordering of the grid cells within the vector is arbitrary, but consistent throughout the entire calculation. The entire set of observations will be grouped into the input matrix $\mathbf{O}$, with observations in rows and grid cells in columns. In the input data used for this study, each observation is 288 (longitude) $\times$ 192 (latitude), for a total of 55296 grid cells. Therefore, in this case, $\mathbf{O}$ has dimension 855×55296.

We will also derive from the input an operator for computing the area-weighted mean of a grid state. We denote this vector by

$$|\lambda\rangle = \frac{1}{S}\sin(|\theta\rangle), \tag{1}$$

where θ is the polar angle (*i.e.*, *colatitude*) of each grid cell, and S is the sum of all the area weights across the entire grid. When defined this way, the global mean temperature for a grid state $|x\rangle$ is $T_g = \langle\lambda\,|x\rangle = \langle x\,|\lambda\rangle$. Similarly, the matrix-vector multiplication $|T_g\rangle = \mathbf{O}\,|\lambda\rangle$ produces a vector of global mean temperature values for the entire input data set.

2.3 Mean response model

Our basic procedure will be to construct a deterministic model for the mean response to global temperature. The residuals between the mean response and the observed temperature fields will be taken as representative of the variability in the ESM and used to construct variability fields that will be added to the mean response to produce the final product.

In principle the mean response could be calculated using any of the emulation techniques described in section 1. For illustrative purposes we will stick with a simple linear pattern scaling using a linear regression variant similar to that described in Mitchell et al. (1999). Using standard least-square regression techniques we compute vectors of coefficients $|w\rangle$ and biases $|b\rangle$ such that the mean response field $|m\rangle$ for temperature T_g is given by

$$|m\rangle\,(T_g) = T_g\,|w\rangle + |b\rangle\,. \tag{2}$$

This formula can be applied to the entire input data set to produce the residual matrix

$$\mathbf{R} = \mathbf{O} - (|T_g\rangle\,\langle w| + |b\rangle)\,, \tag{3}$$

which will be used to construct the variability model. Conversely, the variability fields generated will be added to the mean response (*i.e.*, the last term of equation (3)) to generate absolute temperature fields.

2.4 Generating variability

The matrix of residuals, $\mathbf{R}$, characterizes the variability in the input data. We deem a generated variability data set to be realistic if it matches the distribution of residual values in each grid cell and the space and time correlation properties of the residuals. Our task, therefore, is to generate a random field with specified distribution and correlation properties.

To capture the time correlation we will make use of the Wiener-Khinchin Theorem (Champeney, 1973, § 5.4). This theorem states that given a function $g(t)$ and its Fourier transform $G(f)$,

$$\mathcal{F}(C(g)) = |G(f)|^2, \tag{4}$$

where $C(g)$ is the time autocorrelation function of $g(t)$, and $\mathcal{F}(C)$ is the Fourier transform of C. The salient feature of equation (4) is that the right-hand side of the equation depends only on the magnitude of G. Therefore, we can generate an alternate function g' by setting $|G'| = |G|$, randomizing the phases of G', and taking the inverse Fourier transform. When g' is constructed this way, the Wiener-Khinchin Theorem guarantees that g and g' will have the same autocorrelation function.

In theory we could use a similar technique to capture the spatial correlation; however, in practice the spherical geometry of the spatial domain makes this difficult. Moreover, it is not just the spatial correlation properties that matter, but also the locations at which spatially correlated phenomena occur. Therefore, we capture spatial correlations by using principal components analysis (PCA) to express the grid state as a linear combination of basis vectors that diagonalize the covariance matrix of the system.

$$|x(t)\rangle = \sum_{i=1}^{L} \phi_i(t)\,|\hat{x}\rangle_i, \tag{5}$$

where

$$\left.\begin{aligned} \langle \hat{x}_i|\,\hat{x}_j\rangle = 0,\\ \mathrm{cov}(\phi_i,\phi_j) = 0, \end{aligned}\right\} \qquad \text{if } i \neq j. \tag{6}$$

The $|\hat{x}\rangle_i$ are called *empirical orthogonal functions* (EOFs) (Kutzbach, 1967). The $\phi_i(t)$ are the *projection coefficients* for the grid state vectors. The second property in equation (6) is of particular interest for this application. Because the covariances of the projection coefficients for different EOFs are zero, we can choose them independently. In particular, when applying the phase randomization procedure described above, we can apply it to each ϕ_i independently because all of the spatial correlation properties of the system have been absorbed into the definition of the EOFs.

The EOFs are computed using singular value decomposition (SVD) (Golub and Van Loan, 1996, § 2.5.3). In practice, we introduce some small modifications to the standard procedure. The first is that we define the zeroth basis vector $|\hat{x}\rangle_0$ to be the global mean operator, normalized to unit magnitude:

$$|\hat{x}\rangle_0 = \frac{|\lambda\rangle}{\sqrt{\langle\lambda\,|\lambda\rangle}}. \tag{7}$$

Table 1. The variability generation algorithm

1. Select and fit the mean response model.
2. Construct residual field $\mathbf{R}$ by subtracting mean response from ESM output.
3. Orthogonalize residuals with respect to EOF-0 (equation (7)).
4. Perform the EOF analysis on the residual field.
5. Drop EOFs with singular values $\sigma < 10^{-8}\sigma_{\text{max}}$.
6. Compute the DFT Φ of the residual field's projection coefficients onto the EOF basis.
7. Compute a new Fourier transform $\Phi^{\star}$ such that $|\Phi| = |\Phi^{\star}|$ and the phases of $\Phi^{\star}$ are chosen randomly, uniformly on the interval $[0, 2\pi)$.
8. Compute the projection coefficients $\phi^{\star}$ of the variability field as the inverse DFT of $\Phi^{\star}$.
9. Compute the variability field as $|x(t)\rangle = \sum_{i=0}^{N} \phi_i^{\star}(t)\, |\hat{x}\rangle_i$.

We force this to be a basis vector by subtracting from each residual vector its projection onto $|\hat{x}\rangle_0$ and performing the SVD on the modified residuals. This procedure guarantees that all of the basis vectors have zero area-weighted global mean, or, equivalently, that all of the variability in the global mean temperature is carried in ϕ_0. This property is useful because it allow us to control how much the generated variability distorts the global properties of the mean response field it is being added to. If $\phi_0(t) = 0$, then the global mean will not be affected at all. On the other hand, if it is desirable to change the global means, perhaps because they were generated by a simple climate model (Hartin et al., 2015; Meinshausen et al., 2011) that produces smoother results than real ESMs, then that can be done by setting ϕ_0 appropriately.

In practice $\mathbf{R}$ will usually be rank deficient or nearly so. Therefore, we drop from the basis set any EOFs for which the corresponding singular values are below a threshold defined by the largest singular value found in the SVD, $\sigma_{\text{threshold}} = 10^{-8}\sigma_{\text{max}}$.

At this point we are ready to apply the Wiener-Khinchin Theorem. We compute the discrete Fourier transform (DFT) of the ϕ from equation (5): $\Phi(f) = F(\phi(t))$. We then compute $\Phi^{\star}(f)$ such that $|\Phi^{\star}| = |\Phi|$, but we choose the phases of $\Phi^{\star}$ to be uniform random deviates on the interval $[0, 2\pi]$. From this we can reconstruct $\phi^{\star}(t)$ as the inverse DFT of $\Phi^{\star}(f)$. Finally, we construct the variability field using equation (5), replacing ϕ with $\phi^{\star}$.

The steps in the variability generation algorithm are summarized in Table 1.

3 Results

To illustrate the algorithm, we have produced four independent variability fields by applying the algorithm to the input data described in section 2.2. Training the emulator (*i.e.*, read-in and analysis of the ESM input) took approximately 143 seconds. Each temperature field took 3–4 seconds to generate.

Figure 1 shows a snapshot in time for each of the variability fields. The spatial coherence of the fields is readily apparent. Some features, such as the one seen in the low-latitude eastern Pacific, appear in all of the frames, with greater or lesser strength, or, in one case, with opposite sign. Other features, such as the cool patch over northern Europe in the third frame, have no apparent analog in the other realizations.

We can get a sense of the behavior of the variability fields over time by looking at the power spectral density of the EOFs (fig. 2). Two trends are immediately apparent. First, the total power present in each EOF decreases dramatically after the first few EOFs (fig. 2). The first 10 components together account for 49% of the total power, and the first 50 components account for 72%. Notwithstanding this observation, the long tail of EOFs make a nontrivial contribution to the result. The last 400 EOFs collectively make up a little over 1% of the total power, and as we shall see below, all of the small-scale variability is contained in these components.

The second observation is that the power spectrum whitens (becomes more uniform across frequencies) considerably (Fig. 3), such that only a few of the most prominent EOFs have any significant periodic signature. One interpretation of this observation is that there are only a few consistently repeatable periodic phenomena represented in the surface temperature data of this ESM. The rest of the variability, although highly structured spatially, does not have a lot of temporal structure. The components with significant periodicity account for roughly a third of the total variability signal. In other words, although periodic oscillations are a prominent component of the variability, most of the variability appears to be of the uncorrelated, interannual sort.

In Figure 4 we show the power spectral density for the first nine EOFs. EOF-1 has power primarily at long periods, indicating a pattern of variability that is largely locked in at the beginning of a run, but which varies from one run to the next. EOFs 2, 3, and 5 show evidence of periodicity on time scales ranging from 3 to 20 years.

Figure 5 visualizes the spatial patterns represented by the first 6 EOFs, and Figure 6 visualizes some of the lower power EOFs. These plots show that the scale of the features gets progressively smaller as the power decreases. For example, in EOF-3 there is a complex of positive and negative associations that spans nearly the entire Pacific Ocean. The features visible in EOF-25 are roughly continental scale, while the features in EOF-50 are about half that size. By EOF-400 the feature size is in the hundreds of kilometers, and the lowest power EOF, EOF-853, shows variations a few grid cells in size.

4 Discussion

4.1 Statistical equivalence to ESM input

The time series produced by this method are designed to match three key statistical properties of the ESM data used to train the emulator:

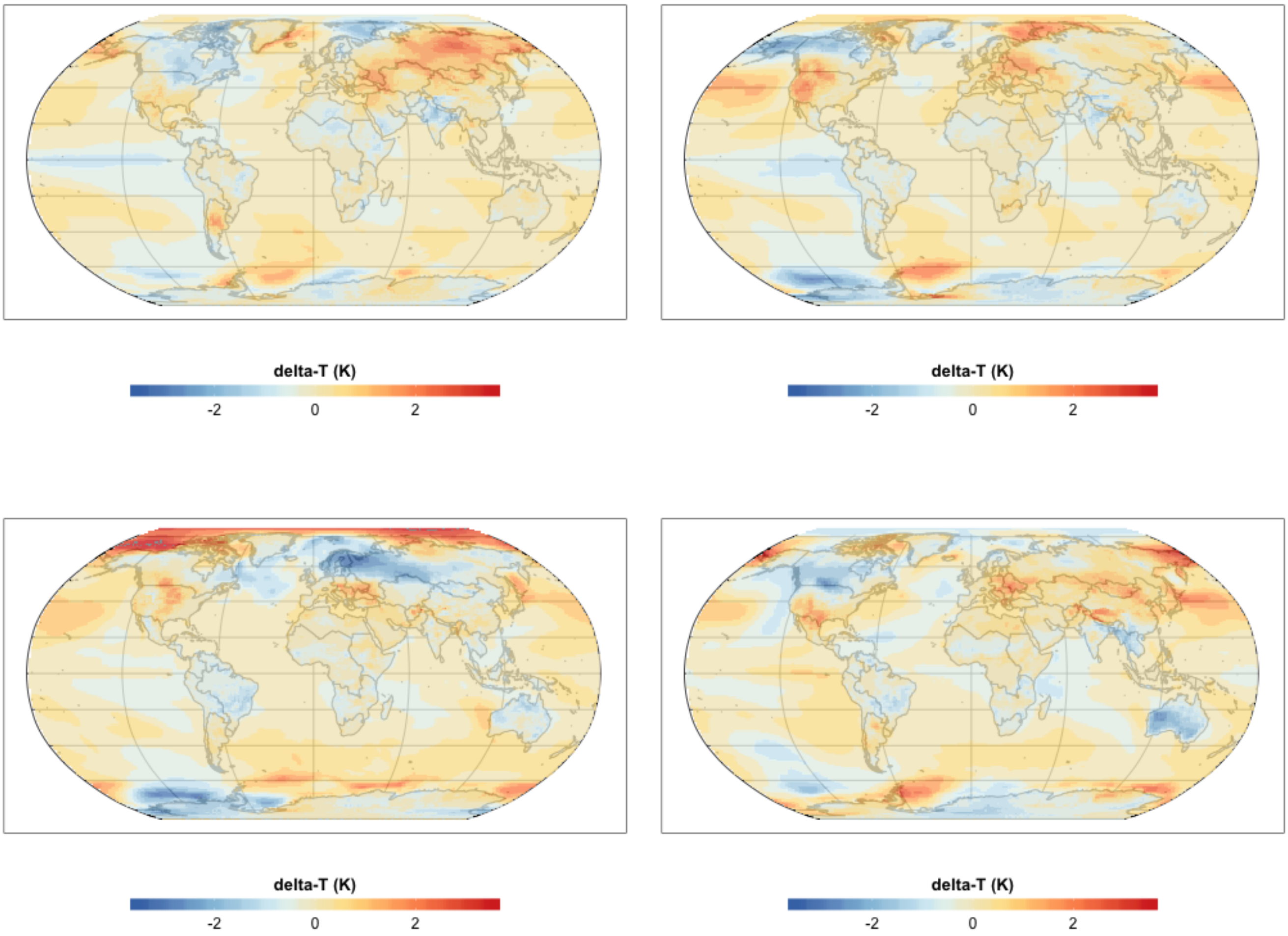


Figure 1. Year 2025 snapshot for variability fields generated using the procedure described in section 2.4. Each field is a different randomly generated realization of the temperature field's departure from the mean response. (sec. 2.3)

1. Distribution of values in a grid cell over time and between realizations.

2. Correlation between values in different grid cells.

3. Time autocorrelation of spatially correlated patterns of grid cells.

In this section we perform a series of statistical tests to verify properties 1 and 2. Property 3 is guaranteed by the Wiener-Khinchin Theorem, and so we do not test it statistically.

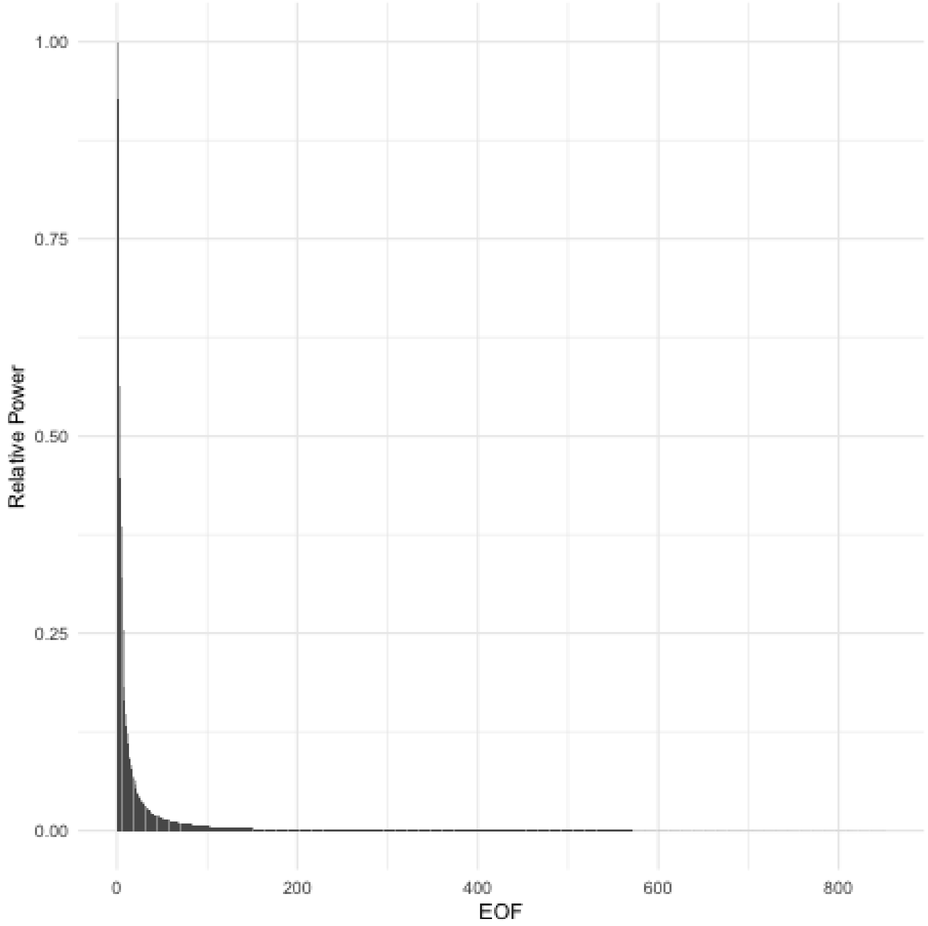


Figure 2. Relative power for each EOF. Roughly half of the total power is contained in the first 10 EOFs. The aggregate power for all EOFs beyond 400 is 1% of the total.

4.1.1 Statistical tests of variability field properties

The generation procedure described in this paper does not strictly guarantee that the generated fields have the desired statistical properties; therefore, we turn to statistical tests of some of the key properties. Testing for the *absence* of an effect is tricky. One cannot simply run a hypothesis test and, seeing a lack of a positive result, conclude that there is no effect. The procedure we have adopted is to focus on tests that can be run in each grid cell (or, in one case, for each pairwise combination of EOFs). We can consider two competing hypotheses:

H1 The statistic being tested is the same in the generated data as in the input data.

H2 The statistic being tested differs in the generated data by some *de minimis* value from the input data.

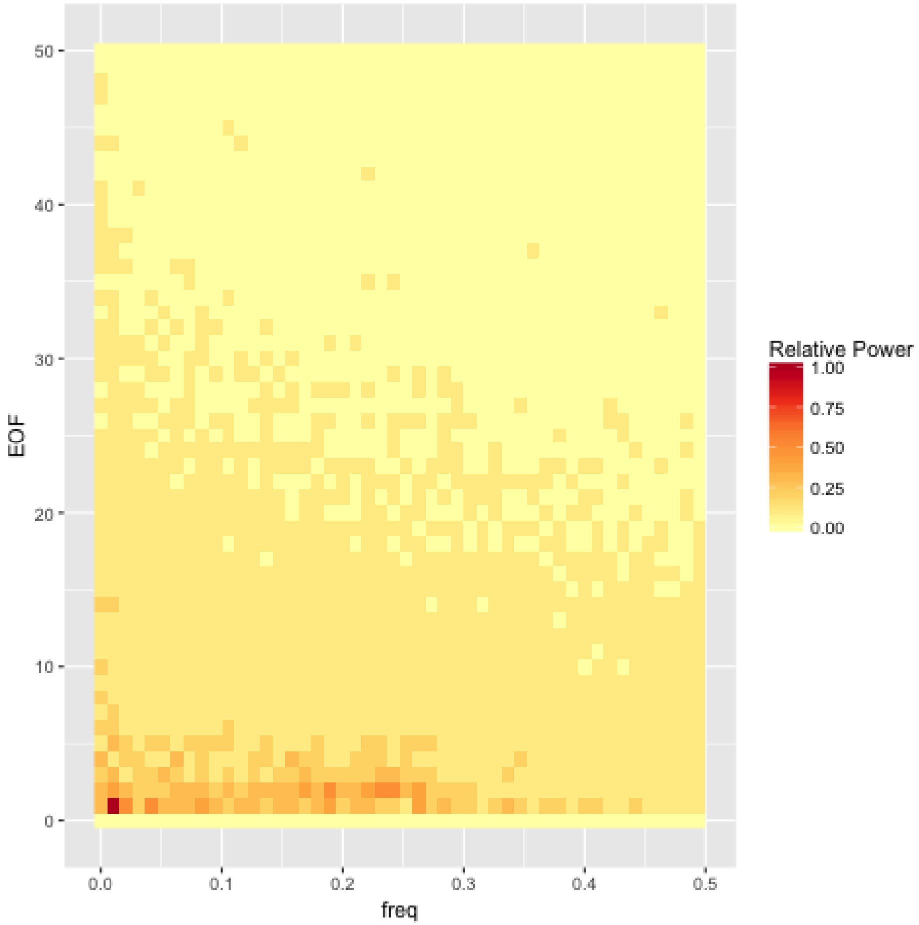


Figure 3. Heat map of power spectral density (PSD) for the first 50 EOFs. The trend of decreasing total power and more uniform spectral density continues for the remaining EOFs beyond EOF-50.

The expected numbers of positive results under these hypotheses are just the p-value (H1) and the power (H2) of the test, each multiplied by the number of tests performed. By observing which of the two hypotheses the actual number of positive results agrees with more closely, we can decide which of the two hypotheses is more likely. The philosophy underlying this procedure is that although we cannot prove that there is *no* statistical difference between the generated and input data, if we can show that an upper bound on the effect size is small enough to be ignorable in practice, then that is sufficient.

All of the statistical tests described in this section were performed on an ensemble of 20 generated fields, each with 95 one-year time steps, for a total of 1900 model outputs in the tests that operated directly on the generated data. For the test that operates on the ϕ values, each temperature grid time series had to be tested separately, for a total of 95 samples per test. In each case the threshold p-value used for the tests was 0.05.

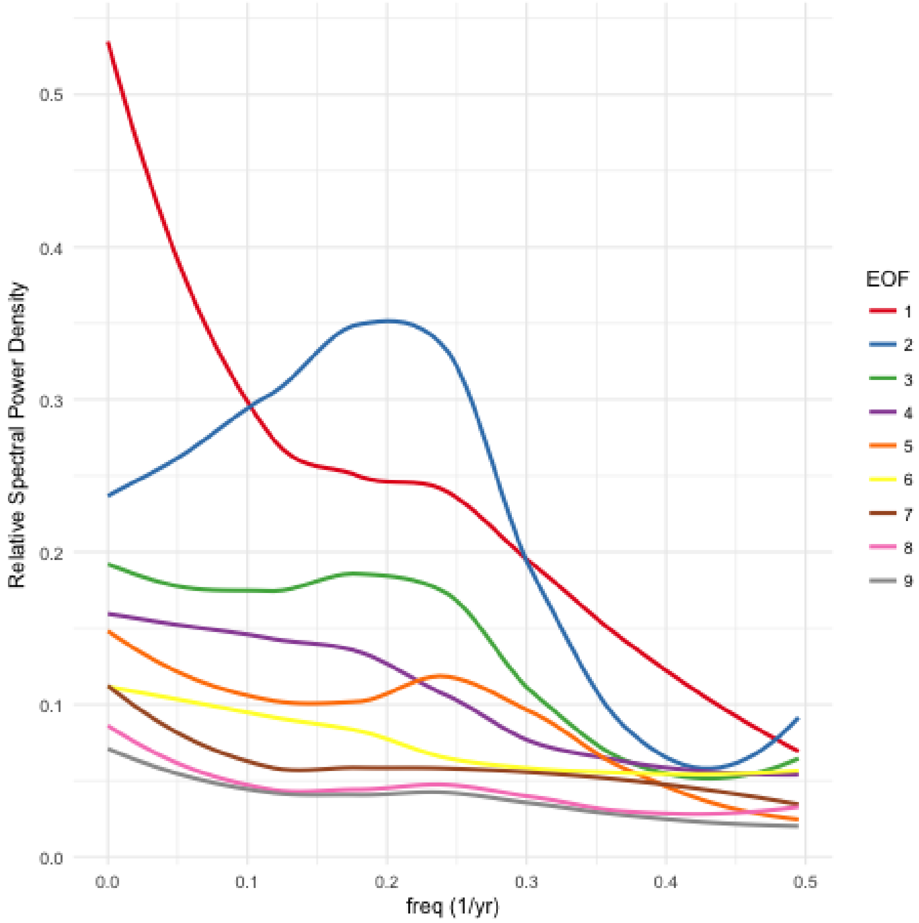


Figure 4. Power spectral density (PSD) for the first 9 EOF basis functions. EOFs 2, 3, and 5 show peaks in the PSD, indicating quasiperiodic behavior. EOFs 1, 3, 4, and 5 also have power at low frequencies.

Table 2. F-test power for several hypothetical percentage differences between input and output variance.

Variance Difference	F-test Power
1%	0.05
2.5%	0.07
5%	0.13
10%	0.37

The first property we will examine is the variance of the distribution of grid cells. We used the F-test of equality of variances to perform this test. In order to be valid, the F-test requires the samples being tested to be normally distributed. We test for this property separately below. Table 2 gives the power (*i.e.*, expected fraction of positive results) for several hypothetical percentage

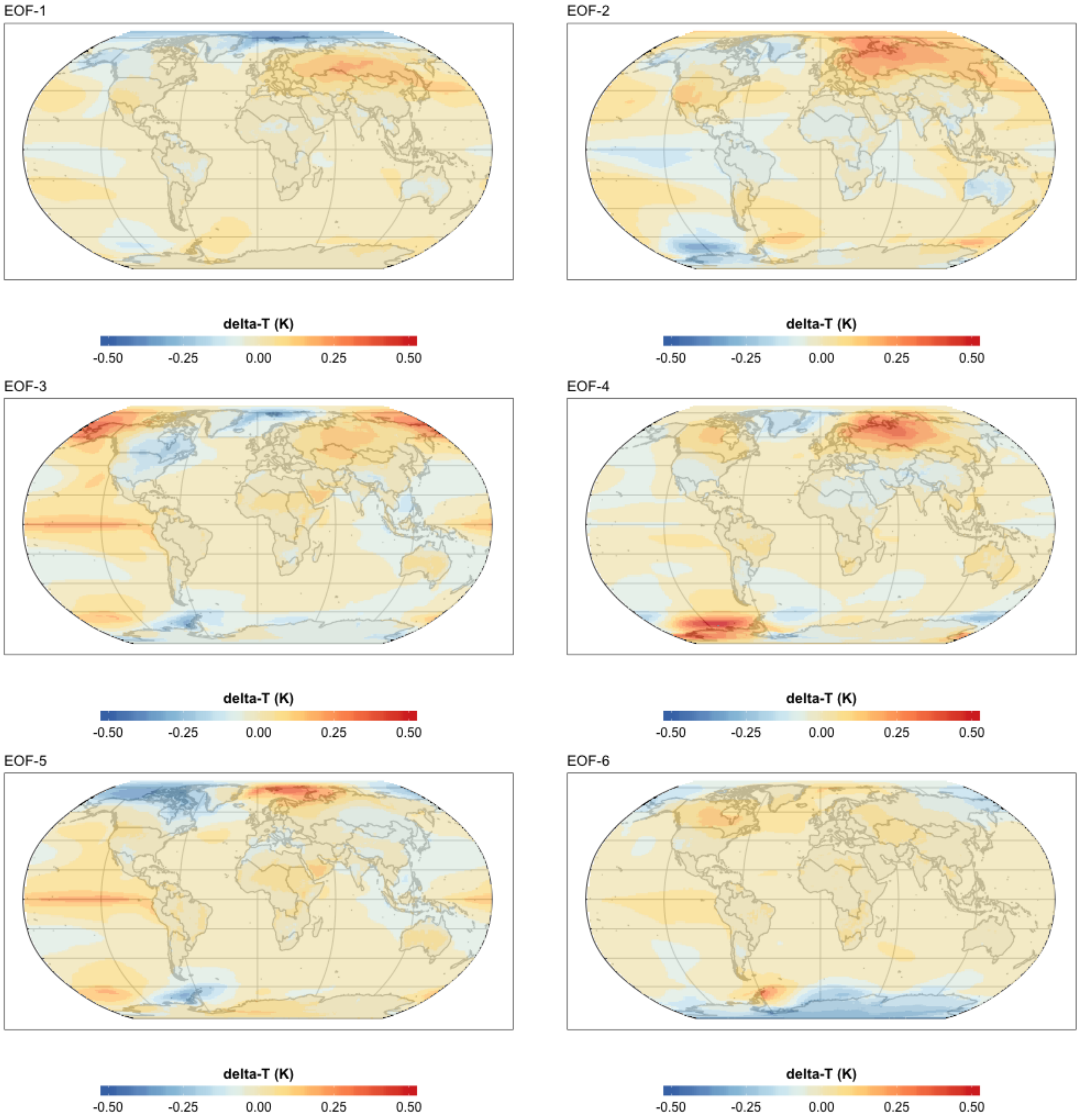
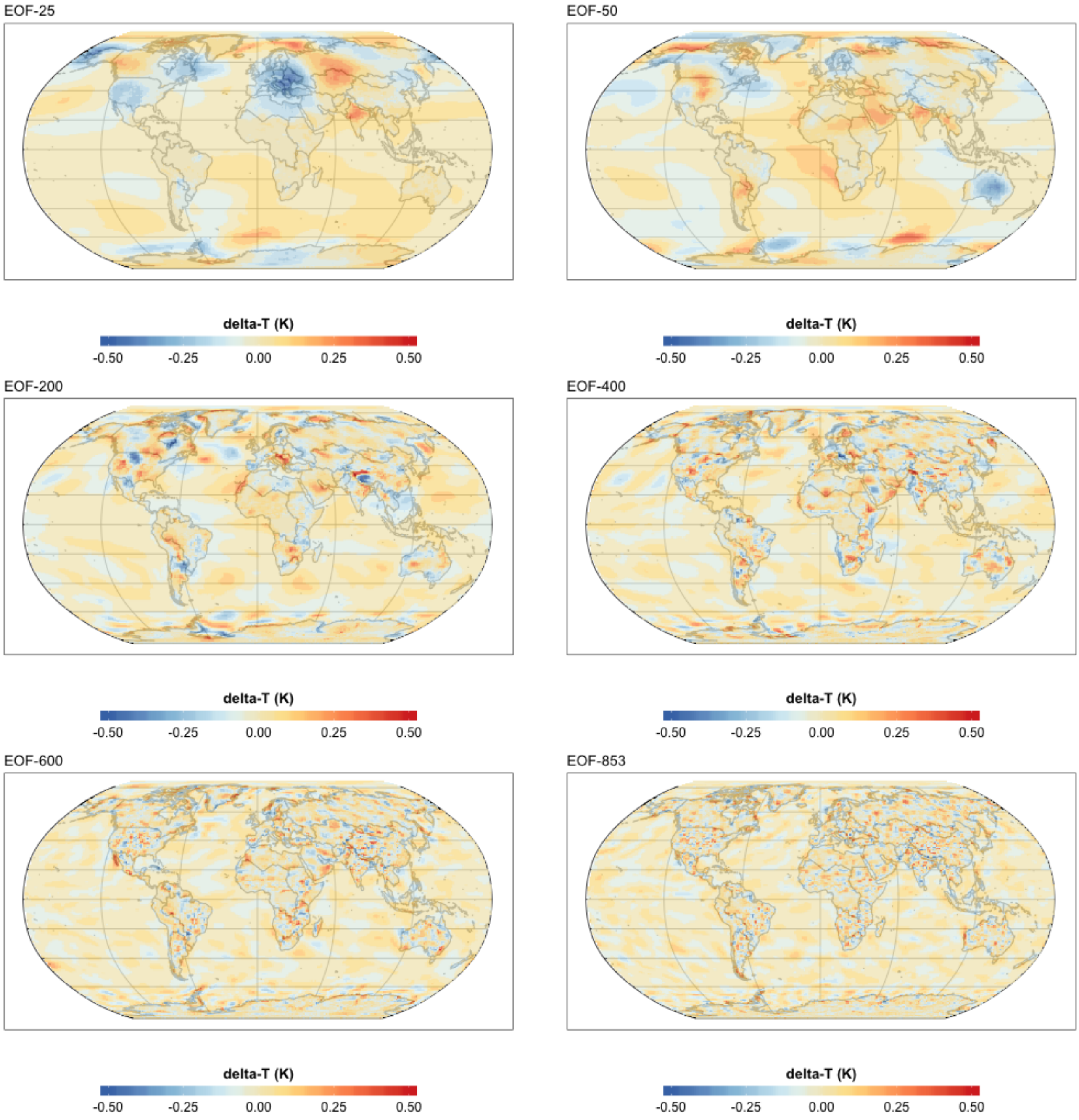


Figure 5. Spatial visualizations of the EOF1-6 basis functions.



EOF-25

EOF-50

EOF-200

EOF-400

EOF-600

EOF-853

Figure 6. Spatial visualizations of higher EOF basis functions. Note the decreasing characteristic scale for basis functions later in the series.

Table 3. Pearson test power for several hypothetical correlation coefficients between ϕ for different EOFs.

Correlation Coefficient	Pearson Test Power
0.01	0.07
0.05	0.59
0.10	0.99

differences in variance between the ESM and generated fields. The actual fraction of positive results was approximately 2×10^{-4}, which is much smaller than the p-value of 0.05.

It may seem surprising that the fraction of positive results was so much smaller than the number expected from the p-value of the tests. This result can be explained by observing that the derivation of the p-value assumes a particular model for H1. Specifically it assumes that the generated data and the reference data (*i.e.*, the ESM input) come from *populations* with exactly equal variance. We cannot observe population variances directly; instead we observe the variances of samples from those populations. The variances of such samples can vary quite a bit from the variance of the underlying population, and so we expect to see some fairly large differences between the variances of input grid cells and the corresponding variances of output grid cells. The F-Distribution tells us just how large we might reasonably expect those discrepancies to be.

Our generated results, on the other hand, are *not* being generated by sampling from a population. Instead, they are generated by a process that seeks to replicate the variances of the reference data exactly. If it were completely successful at doing so, then all of the variances would be identical to their counterparts in the reference set, and there would be precisely zero positive results. In actuality, there are some slight discrepancies, but these are much smaller than the ones assumed in the formulation of H1. Therefore, we see many fewer positive results than would be expected based on the p-value used in the tests.

Our second test concerns the covariance between grid cells. Testing for equal, nonzero covariances directly is challenging, but we can transform the results into a form that is more readily testable. Starting from equation (5) we can show that for two grid cells x_m and x_n

$$\mathrm{cov}(x_m, x_n) = \sum_i \mathrm{var}(\phi_i)\hat{x}_{im}\hat{x}_{in} + \sum_{i \neq j} \mathrm{cov}(\phi_i, \phi_j)\hat{x}_{im}\hat{x}_{jn}, \tag{8}$$

where $\hat{x}_{im}$ is the mth component of $|\hat{x}\rangle_i$. The corresponding expression for the generated data is the same, except that the ϕ are replaced by $\phi^\star$. For the input ESM data, the construction of the EOFs guarantees that $\mathrm{cov}(\phi_i, \phi_j) = 0$, when averaged over the input data. Thus, the grid cell covariances of the generated data will match those of the ESM data if, averaged over runs of the generator:

$$\mathrm{var}(\phi_i^\star) = \mathrm{var}(\phi_i) \qquad \text{for all } i\text{, and} \tag{9}$$

$$\mathrm{cov}(\phi_i^\star, \phi_j^\star) = 0 \qquad \text{for all } i \neq j. \tag{10}$$

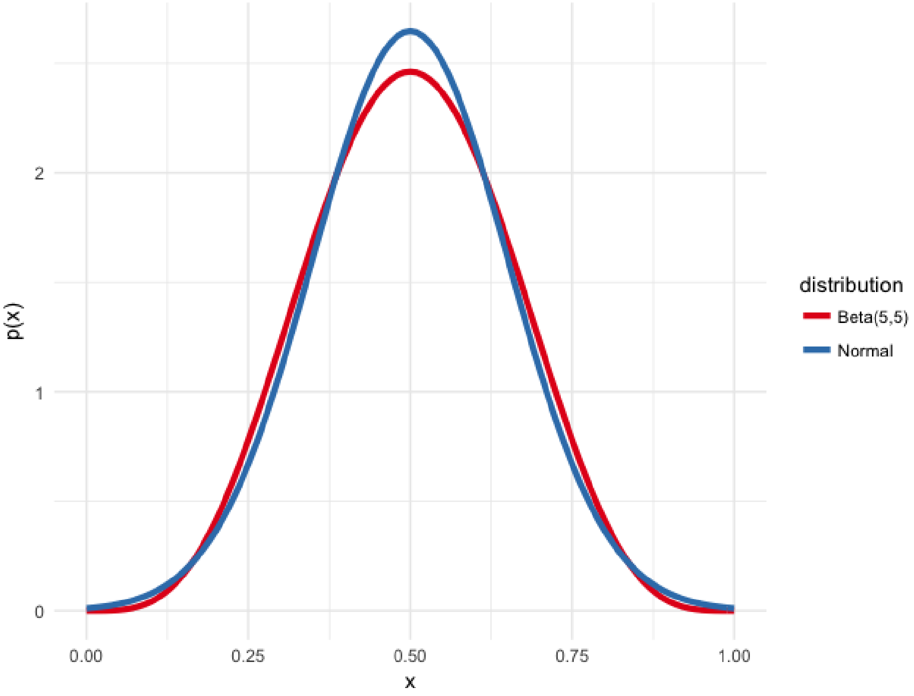


Figure 7. Comparison of the Beta(5,5) distribution and a Normal distribution with equal variance.

The first of these two conditions is guaranteed by the generation procedure. Parseval's Theorem (Champeney, 1973, appendix E) states that for each of the ϕ_i (and likewise for the $\phi_i^\star$),

$$\sum_{t=1}^{N_t} (\phi_i(t))^2 = \sum_{k=1}^{N_t} |\mathcal{F}_k(\phi_i)|^2 . \tag{11}$$

Since our procedure ensures $|\mathcal{F}_k(\phi_i^\star)| = |\mathcal{F}_k(\phi_i)|$, this guarantees that the condition in equation (9) holds.

To test the condition in equation (10) we used Pearson's correlation test. Table 3 gives the power of the test for various correlation coefficients for the alternative hypothesis. The actual fraction of positive tests, over the pairwise combinations of EOFs, was 0.05, or roughly what we would expect from the p-value used in the test. From these observations we can conclude that the upper bound on possible correlation coefficients between the ϕ is somewhere between 0.01 and 0.05.

The final statistical test concerns whether the generated residuals are normally distributed. Apart from being necessary to ensure the validity of the F-tests above, a normal distribution is desirable per se because we expect the temperature residuals to be normally distributed. This test is more challenging to perform than the rest because there is no obvious way to define an effect size to use in calculating the power. Instead, we must determine a reasonable nonnormal distribution to use as the benchmark for deviations from normality.

To arrive at such a distribution, consider how the generated residual fields are calculated. The value x of the residual temperature in each grid cell is produced by summing over all EOFs and all Fourier components. Since the phases of the Fourier components are chosen randomly, this amounts to a sum over uniform random deviates, which by the Central Limit Theorem will be asymptotically normally distributed. Any deviations from normality will be due to having insufficient terms in the sum to reach that asymptotic behavior. Such a distribution would appear truncated compared to the normal distribution, since the sum of uniform random deviates has hard minimum and maximum values. The Beta distribution, $B(n_1, n_2)$ also has these properties. When $n_1 = n_2 = n$, the distribution is symmetric and approaches a Normal distribution as n increases. We adopted the $B(5,5)$ distribution, shown in Figure 7, as our representative distribution for a *de minimis* effect size.

We used the Shapiro-Wilk Test of Normality to evaluate the normality of the grid cell distribution. For this sample size, the power of the test for distinguishing between a $B(5,5)$ and a Normal distribution is 0.998. The actual fraction of grid cells that showed a positive result was 0.06, indicating that if there is any nonnormality, it is almost certainly smaller than the difference between a normal distribution and a $B(5,5)$ distribution.

4.1.2 Commentary on statistical properties

Property 3 deserves additional comment because it is explicitly *not* equivalent to matching the time autocorrelation function of individual grid cells. We chose to focus on autocorrelation of spatial patterns rather than on grid cells because the only way to preserve the autocorrelation of grid cells would be to force a constant phase difference between EOFs. This assumption doesn't seem particularly realistic and isn't supported by the input data. Limiting the treatment of time autocorrelation to the EOFs ensures that to the extent that EOFs represent physical phenomena they occur with the right frequency spectra, while not overly constraining the phase relationships between modes.

The properties enumerated above ensure that, when using the generated data to drive an ensemble of downstream models and compute statistics on those results, the scale of the fluctuations produced, their spatial location and extent, and their periodic character, if any, will be faithfully reproduced, allowing reliable calculations of variance in outcomes, return times of extremes, and regional differences in impacts. Therefore, we expect a technique like this to be invaluable for studies of the contribution of variability to uncertainty in climate effects and feedbacks.

Supporting such uncertainty studies was our primary purpose in developing this tool, but the analysis in section 3 suggests additional possibilities. A byproduct of the procedure to generate variability fields is that we develop quite a few statistics that could be used to characterize the ESM used to train the emulator. Thus, the training stage of the emulation procedure could also function as a diagnostic package for ESMs. For example, the high power at low frequencies for the first 10–15 EOFs (Fig. 3) was unexpected and might be of interest for further study.

4.2 Overfitting the mean response

There is one important pitfall to watch out for when using this method to learn the behavior of an ESM; viz., one must take care not to allow the mean response model to overfit the ESM data. The more complex the model, the greater the danger of overfitting, but even simple models like the linear regression used here can overfit. Consider EOF-1 and its power spectrum, depicted in figure 4. The power spectrum's strong peak at $f = 0$ means that the coefficient ϕ_1 of the component is nearly constant within a single run of ESM data. Therefore, if we were to train the model on just a single run (*i.e.*, a single realization of a single scenario), this component would be absorbed into the mean response, causing it to be reproduced identically in all generated temperature fields. In fact, this is precisely what happened in early versions of this work, where we trained the emulator on a single ESM run. EOF-1 only began to appear in the variability fields once we expanded the input data to include the full suite of CESM(CAM5) runs from CMIP5.

Therefore, it is essential to include enough independent ESM runs in the training data to ensure that the mean response model will not capture fluctuations that are idiosyncratic to a particular run. Exactly how many runs are needed will depend on the complexity of the mean field response model. For a relatively simple model, such as the linear model used in this paper, as few as three independent runs (i.e., one more than the number of parameters per grid cell) should provide reasonable protection against absorbing variability features into the mean response model. Conversely, mean response models with many parameters per grid cell would require more independent inputs. In case of doubt, cross-validation should be used to diagnose possible overfitting. Along similar lines, the input data should include runs for scenarios that span the entire range of future scenarios that the system will be used to emulate. This practice ensures that the mean response model will not be called upon to extrapolate beyond the range of conditions it was trained on.

4.3 Assumptions

As with most emulation schemes, this one makes certain assumptions about the models it is trying to emulate. The most important assumption is that the ESM outputs can be linearly separated into a temperature-dependent component (what we've been calling the "mean field response") and a time-dependent component (the "variability"). Notably, we assume that the temperature response is independent of the temperature history. This assumption, though common in emulator studies, is dubious. The assumption can be partially negated by including additional predictor variables in the mean field model (*e.g.* Joshi et al., 2013; MacMartin and Kravitz, 2016). At the same time, the second assumption implies that the internal dynamics of the ESM are unaffected by the specifics of the external forcing, which is certainly debatable.

A related assumption is the assumption of stationarity. The variability fields produced by this method have stationary statistical properties. Some research has suggested that the variability is likely to change with increasing global mean temperature (Murray and Ebi, 2012). This sort of phenomenon could be added to our method by introducing a global mean temperature-dependent scale factor. Such a factor would be applied in between steps 8 and 9 in Table 1.

5 Conclusions

Having a computationally efficient method for generating realizations of future climate pathways is a key enabler for research into uncertainties in climate impacts. In order to be fit for this purpose, a proposed method must produce data with statistical properties that are similar to those of Earth System Models, which are currently the state of the art in projecting future climate states.

In the preceding sections we have described such a method, and we have shown that it reproduces key statistical properties of the Earth System Model on which it was trained. Specifically, it produces equivalent distributions of residuals to the mean field response and equivalent space and time correlation structure. The method is computationally efficient, requiring under 10 minutes to train on the input data set used for the results presented here. Once training is complete, generating temperature fields takes just a few seconds per field generated.

As a result, we believe the method will be extremely useful for the impacts studies it was designed to support. Currently, the method is limited to producing temperature only, and at annual resolution. However, we believe that the method can be readily extended to other climate variables and to shorter time scales. These extensions will be the subject of follow-up work.

Code and data availability. Software implementing this technique is available as an R package released under the GNU General Public License. Full source can be found in the project's GitHub repository (https://github.com/JGCRI/fldgen). Release version 1.0.0 of the package was used for all of the work in this paper.

The data and analysis code for the results presented in this paper are archived at https://doi.org/10.5281/zenodo.1183641.

Author contributions. Link designed the algorithm, developed the fldgen package and performed the analysis of the results. Lynch ran early versions of the algorithm and performed analysis on those results. Snyder performed the theoretical analysis of the algorithm's properties, which informed the statistical analysis. Hartin, Kravitz, and Bond-Lamberty advised the project and provided revisions and feedback for early drafts of the paper. Link prepared the manuscript with contributions from all coauthors.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This research is based on work supported by the US Department of Energy, Office of Science, Integrated Assessment Research Program. The Pacific Northwest National Laboratory is operated for DOE by Battelle Memorial Institute under contract DE-AC05-76RL01830.

References

Alexeeff, S. E., Nychka, D., Sain, S. R., and Tebaldi, C.: Emulating mean patterns and variability of temperature across and within scenarios in anthropogenic climate change experiments, Climatic Change, https://doi.org/10.1007/s10584-016-1809-8, https://doi.org/10.1007/s10584-016-1809-8, 2016.

Bodman, R. W. and Jones, R. N.: Bayesian estimation of climate sensitivity using observationally constrained simple climate models, Wiley Interdisciplinary Reviews: Climate Change, 7, 461–473, https://doi.org/10.1002/wcc.397, http://dx.doi.org/10.1002/wcc.397, 2016.

Castruccio, S. and Stein, M.: Global space-time models for climate ensembles, The Annals of Applied Statistics, 7, 1593–1611, 2013.

Champeney, D. C.: Fourier Transforms and Their Physical Applications, Academic Press, New York, 1973.

Falloon, P., Challinor, A., Dessai, S., Hoang, L., Johnson, J., and Koehler, A.-K.: Ensembles and uncertainty in climate change impacts, Frontiers in Environmental Science, 2, 33, http://journal.frontiersin.org/article/10.3389/fenvs.2014.00033, 2014.

Golub, G. H. and Van Loan, C. F.: Matrix Computations, JHU Press, Baltimore, MD, 3 edn., 1996.

Greenough, G., McGeehin, M., Bernard, S. M., Trtanj, J., Riad, J., and Engelberg, D.: The potential impacts of climate variability and change on health impacts of extreme weather events in the United States., Environmental Health Perspectives, 109, 191 – 198, 2001.

Hartin, C. A., Patel, P., Schwarber, A., Link, R. P., and Bond-Lamberty, B. P.: A simple object-oriented and open-source model for scientific and policy analyses of the global climate system – Hector v1.0, Geoscientific Model Development, 8, 939–955, https://doi.org/10.5194/gmd-8-939-2015, http://www.geosci-model-dev.net/8/939/2015/, 2015.

Holden, P. B. and Edwards, N. R.: Dimensionally reduced emulation of an AOGCM for application to integrated assessment modelling, Geophysical Research Letters, 37, https://doi.org/10.1029/2010GL045137, http://dx.doi.org/10.1029/2010GL045137, 2010.

Joshi, M., Lambert, F., and Webb, M.: An explanation for the difference between twentieth and twenty-first century land–sea warming ratio in climate models, Climate dynamics, 41, 1853–1869, 2013.

Kay, J., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., Arblaster, J., Bates, S., Danabasoglu, G., Edwards, J., et al.: The Community Earth System Model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability, Bulletin of the American Meteorological Society, 96, 1333–1349, 2015.

Kutzbach, J. E.: Empirical eigenvectors of sea-level pressure, surface temperature and precipitation complexes over North America, Journal of Applied Meteorology, 6, 791–802, 1967.

MacMartin, D. G. and Kravitz, B.: Dynamic climate emulators for solar geoengineering, Atmospheric Chemistry and Physics, 16, 15 789–15 799, https://doi.org/10.5194/acp-16-15789-2016, http://www.atmos-chem-phys.net/16/15789/2016/, 2016.

Meinshausen, M., Raper, S. C. B., and Wigley, T. M. L.: Emulating coupled atmosphere-ocean and carbon cycle models with a simpler model, MAGICC6 –Part 1: Model description and calibration, Atmos. Chem. Phys., 11, 1417–1456, https://doi.org/10.5194/acp-11-1417-2011, http://www.atmos-chem-phys.net/11/1417/2011/, 2011.

Mitchell, J., Johns, T. C., Eagles, M., Ingram, W. J., and Davis, R. A.: Towards the construction of climate change scenarios, Climatic Change, 41, 547–581, 1999.

Mitchell, T. D.: Pattern Scaling: An Examination of the Accuracy of the Technique for Describing Future Climates, Climatic Change, 60, 217–242, https://doi.org/10.1023/A:1026035305597, http://dx.doi.org/10.1023/A:1026035305597, 2003.

Murphy, J. M., Sexton, D. M. H., Barnett, D. N., Jones, G. S., Webb, M. J., Collins, M., and Stainforth, D. A.: Quantification of modelling uncertainties in a large ensemble of climate change simulations, Nature, 430, 768–772, http://dx.doi.org/10.1038/nature02771, 2004.

Murray, V. and Ebi, K. L.: IPCC special report on managing the risks of extreme events and disasters to advance climate change adaptation (SREX), 2012.

Neelin, J. D., Bracco, A., Luo, H., McWilliams, J. C., and Meyerson, J. E.: Considerations for parameter optimization and sensitivity in climate models, Proceedings of the National Academy of Sciences, 107, 21 349–21 354, 2010.

Osborn, T. J., Wallace, C. J., Harris, I. C., and Melvin, T. M.: Pattern scaling using ClimGen: monthly-resolution future climate scenarios including changes in the variability of precipitation, Climatic Change, pp. 1–17, https://doi.org/10.1007/s10584-015-1509-9, http://dx.doi.org/10.1007/s10584-015-1509-9, 2015.

Rasmussen, D. J., Meinshausen, M., and Kopp, R. E.: Probability-Weighted Ensembles of U.S. County-Level Climate Projections for Climate Risk Analysis, Journal of Applied Meteorology and Climatology, 55, 2301–2322, https://doi.org/10.1175/JAMC-D-15-0302.1, http://dx.doi.org/10.1175/JAMC-D-15-0302.1, 2016.

Sanderson, B. M., Oleson, K. W., Strand, W. G., Lehner, F., and O'Neill, B. C.: A new ensemble of GCM simulations to assess avoided impacts in a climate mitigation scenario, Climatic Change, pp. 1–16, https://doi.org/10.1007/s10584-015-1567-z, http://dx.doi.org/10.1007/s10584-015-1567-z, 2015.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An overview of CMIP5 and the experiment design, Bulletin of the American Meteorological Society, 93, 485–498, 2012.

Tebaldi, C. and Arblaster, J. M.: Pattern scaling: Its strengths and limitations, and an update on the latest model simulations, Climatic Change, 122, 459–471, https://doi.org/10.1007/s10584-013-1032-9, http://dx.doi.org/10.1007/s10584-013-1032-9, 2014.