Fldgen v1.0: ~~A Computationally Efficient~~ An Emulator with Internal Variability and Space-Time Correlation for Earth System Models

Robert Link[1], Abigail Snyder[1], Cary Lynch[2], Corinne Hartin[1], Ben Kravitz[3,4], and Ben Bond-Lamberty[1]

[1]Pacific Northwest National Laboratory, Joint Global Change Research Institute, 5825 University Research Ct., College Park, MD, USA
[2]Connecticut Department of Energy and Environmental Protection, 10 Franklin Square New Britain, CT, USA
[3]Department of Earth and Atmospheric Sciences, Indiana University, 1001 E. 10th St., Bloomington, IN, USA
[4]Atmospheric Sciences and Global Change Division, Pacific Northwest National Laboratory, 902 Battelle Boulevard, Richland, WA, USA

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

Authors' responses to reviewers' comments

Anonymous Reviewer #2

> - The size of the vector b in figure 1 should be (55296 x 1), shouldn't it? The reviewers' responses suggest so, and working through the calculation and text myself, this is how I understand it.
>
> An aside: I know the software is written in R, but the broadcasting explaination in section 2.1 invokes numpy. Technically, numpy broadcasting would only work here if the array b was of size (55296). The text implies that the array size is (55296 x 1), which is fundamentally different in numpy. As a Python native this confused me for a while, but the authors have explained their notation and broadcasting choice in section 2.1, so I am not too concerned about this.

Both of these observations are correct. I have fixed Figure 1 accordingly, including dropping the second index on b. However, b is still depicted graphically as a column vector, since that seems to be how most people regard "vectors" of unspecified shape.

> - page 7, line 17: 3 to 20 years. I think 3 to 5 years is a bit more apparent; I don't see any noticeable periodicity on frequencies < 0.2. The captions to figures 5 and 6 also say 3-5 years.

The skewed shape of the PSD for EOF-2 makes it a little nebulous what the lower bounds of the frequencies in that mode are. I have changed the frequency range in the description to 3–5 years to coincide more closely with the peak of the PSD and for consistency with the figure captions.

Thank you for your helpful suggestions.

Anonymous Reviewer #3

The title is not very informative - all emulators are computationally efficient, so what marks this one out? A title that refers to internal variability might be useful, and would help to draw the attention of target end-users?

Specific examples of cases when impacts are strongly dependent upon internal variability might be useful to make it crystal clear what this is doing and why it is needed. Even just pulling in the headline result from Ray "Climate variation explains a third of global crop yield variability" would help non-specialist readers a lot I think.

These are both excellent suggestions, which I have implemented. Thank you.

I had similar thoughts as reviewer two re using different RCPs used to train the emulator, and I didn't find the response fully convincing – some tests should be possible here? In particular, I found myself wondering whether EOF1 reflected responses to different RCPs. Did you check whether EOF1 was similar in an emulator trained on a few simulations with a single RCP (and didn't disappear, as when you trained on a single simulation)? Why not compare the variance of the full emulator (i.e. trained on all 9 simulations) separately with the RCP2.6 simulations and with the RCP8.5 simulations to quantify any scenario bias? In any event, some discussion of this should be included in the text as I suspect many people will question it. Repeating from Reviewer Two "Section 4.2 got me thinking that as the model is trained on the RCP outputs, is there any difference in the results when taking just the set of realisations from RCP2.6 and RCP8.5? Certainly across ESMs, the variance across models increases with increasing global mean temperature. It would therefore not be correct to use a variability model that is trained on RCP8.5 for low forcing scenarios or those with a peak and decline. I note the authors address this in section 4.3, but I wonder if they have tested this."

My analysis of this point wound up being rather long, so I have placed it in our public code and data archive. It can be downloaded from https://zenodo.org/record/2586040/files/cc-analysis.nb.html?download=1 The summary is that the mean response model trained on the ensemble members from a single RCP is practically indistinguishable from the mean response model trained on an equivalent number of ensemble members from different RCPs. I have added a new subsction 3.4, which compares these mean field models and directs readers to the archive for further tests.

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 ESM output 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 applications that require them. 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 be useful in the study of climate variability and its contribution to unertainties in the interactions between human and earth systems.

Copyright statement. TEXT

1 Introduction

There are a variety of scientific applications that use data from future climate scenarios as input. Examples include crop and agricultural productivity models (Rosenzweig et al., 2014; Elliott et al., 2014; Nelson et al., 2014), water and hydrology models (Cui et al., 2018; Voisin et al., 2017), energy models (Turner et al., 2017), and global human systems models (Akhtar et al., 2013; Calvin and Bond-Lamberty, 2018). Earth System Models (ESMs) are the gold standard for producing these future projections of climate change; however, running ESMs is difficult and costly. As a result, most users of ESM data 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 produced 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 of scenarios 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 they might need many realizations of one or more future climate scenarios.

Examples of research areas for which archival runs might be insufficient include uncertainty studies, in which the multiple realizations are used to compute a statistical distribution of outcomes in the downstream model (Murphy et al., 2004; Falloon et al., 2014; Sanderson et al., 2015; Bodman and Jones, 2016; Rasmussen et al., 2016). Studying tail risk (*i.e.*, the effects of climate variables assuming values in the tails of their distribution, which by definition occurs infrequently in any single scenario run) is another example (Greenough et al., 2001), and studying sensitivity to climate variability is a third (Kay et al., 2015).

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." Many of the applications described above are sensitive to climate variability~~(e.g. Ray et al., 2015), so capturing it~~. For example, Ray et al. (2015) found that "Globally, climate variability accounts for roughly a third (~~32--~~39%) of the observed yield variability" in agricultural crops. Therefore, capturing this variability in emulators is crucial to understanding the behavior of and uncertainties in these ~~models~~applications.

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 underlined bold symbols (*e.g.* $\underline{\mathbf{R}}$) to refer to matrices. Ordinary bold symbols are used for vectors (*e.g.* $\boldsymbol{x}$). When it is necessary to distinguish between column and row vectors, the latter will be marked as the transpose of a column vector (*e.g.* $\boldsymbol{x}^\top$). These vectors represent collections of scalar quantities that bear some relationship to each other in time or space. Because of this, the same variable can appear in both vector and scalar variants, with the vector decoration (or lack thereof) indicating which is meant. For example, T_g is the global mean temperature, a scalar, while $\boldsymbol{T_g}$ is a vector representing a sequence of global mean temperatures.

Occasionally we will add a matrix and a vector; *e.g.*, $\underline{\mathbf{B}} = \underline{\mathbf{A}} + \boldsymbol{x}$. This should be interpreted to mean that the vector $\boldsymbol{x}$ is to be added to each row of the matrix $\underline{\mathbf{A}}$. Therefore, the length of $\boldsymbol{x}$ must be equal to the number of columns in $\underline{\mathbf{A}}$. 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 global temperature states.

To keep clear the distinction between the data produced by the emulator and the ESM data used to train the emulator, we will refer to the ESM data as "synthetic measurements" (when referring to the data as a whole) or "cases" (when referring to individual frames in the data), while the terms "results" and "model output" will be reserved for the data produced by the emulator.

Throughout the discussion, we will treat each temperature state 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 synthetic measurements will be grouped into the input matrix $\underline{\mathbf{O}}$, with the cases in rows and grid cells in columns. In the input data used for this study, each case is 288 (longitude) $\times$ 192 (latitude), for a total of 55296 grid cells. Therefore, in this case, $\underline{\mathbf{O}}$ 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

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

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 $\boldsymbol{x}$ is $T_g = \boldsymbol{\lambda}^\top \boldsymbol{x} = \boldsymbol{x}^\top \boldsymbol{\lambda}$. Similarly, the matrix-vector multiplication $\boldsymbol{T_g} = \underline{\mathbf{O}}\boldsymbol{\lambda}$ 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 synthetic 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 weights $\boldsymbol{w}$ and biases $\boldsymbol{b}$ (each of these vectors has length equal to the number of grid cells) such that the mean response field $\boldsymbol{m}$ for global mean temperature T_g is given by

$$\boldsymbol{m}(T_g) = T_g\boldsymbol{w} + \boldsymbol{b}. \tag{2}$$

This formula can be applied to the entire input data set, with $T_g\boldsymbol{w}$ becoming the outer product $\boldsymbol{T_g}\boldsymbol{w}^\top$ to produce the residual matrix

$$\underline{\mathbf{R}} = \underline{\mathbf{O}} - \left(\boldsymbol{T_g}\boldsymbol{w}^\top + \boldsymbol{b}\right), \tag{3}$$

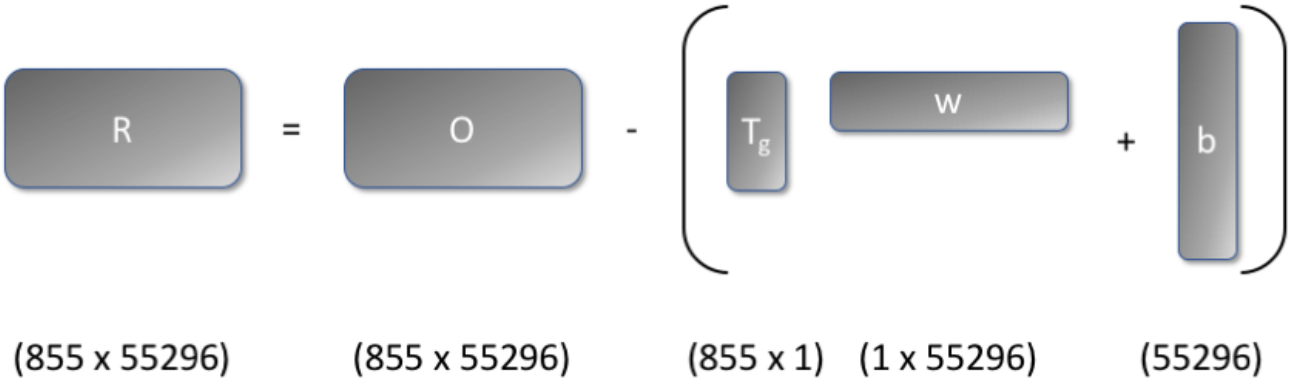

Figure 1. Schematic of the residual calculation showing the shapes of the matrices involved. The result of the outer product $\boldsymbol{T_g w}^\top$ is an 855×55296 matrix. The vector $\boldsymbol{b}$ is added to this matrix using the broadcast convention described in section 2.1

which will be used to construct the variability model. This calculation is shown schematically in Figure 1. 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, $\underline{\mathbf{R}}$, 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 magnitudes of the elements of G, not their phases (recall that the results of a Fourier transform are complex numbers with both magnitude and phase). Therefore, we can generate an alternate function g' by setting $|G'| = |G|$, selecting the phases of G' at random, 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.

$$\boldsymbol{x}(t) = \sum_{i=1}^{L} \phi_i(t) \hat{\boldsymbol{x}}_i, \tag{5}$$

where

$$\left.\begin{aligned} \hat{\boldsymbol{x}}_i^\top \hat{\boldsymbol{x}}_j = 0, \\ \operatorname{cov}(\phi_i, \phi_j) = 0, \end{aligned}\right\} \quad \text{if } i \neq j. \tag{6}$$

The $\hat{\boldsymbol{x}}_i$ are called *empirical orthogonal functions* (EOFs) (Kutzbach, 1967) and are computed using singular value decomposition (SVD) (Golub and Van Loan, 1996, § 2.5.3). 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.

In practice, it is convenient to force all of the basis vectors except for one to have area-weighted global means of zero, so that all of the variability in the global mean is carried by a single component. This property is useful because it allows us to control how much the generated variability distorts the global properties of the mean response field it is being added to. To accomplish this, we introduce a small modification to the EOF decomposition procedure. We define the zeroth basis vector $\hat{\boldsymbol{x}}_0$ to be the global mean operator, normalized to unit magnitude:

$$\hat{\boldsymbol{x}}_0 = \frac{\boldsymbol{\lambda}}{\sqrt{\boldsymbol{\lambda}^\top \boldsymbol{\lambda}}}. \tag{7}$$

We force $\hat{\boldsymbol{x}}_0$ to be a basis vector by subtracting from each residual vector its projection onto $\hat{\boldsymbol{x}}_0$ and performing the SVD on the modified residuals. This procedure forces all of the basis vectors to be orthogonal to $\hat{\boldsymbol{x}}_0$. Since this vector is proportional to the global mean operator $\boldsymbol{\lambda}$, this orthogonality property guarantees that all of the other basis vectors will have zero global mean. Therefore, if $\phi_0(t) = 0$, then the global means of the mean response fields will be unaffected when the generated residual fields are added. 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.

The typical use of PCA in many fields, including climate modeling, is for dimensionality reduction. In such applications the next step after computing the EOFs would be to identify and keep a small set of EOFs that capture the majority of the variability and to throw away the rest. In this case, dimensionality reduction is *not* our goal. Rather, we have used the EOF decomposition only to separate the residual field into components that are uncorrelated over time. Therefore, we keep the full set of EOFs and their projection coefficients. The sole exception is for components for which the singular values produced by the SVD procedure are very small. There are generally 1 or 2 such components, and keeping them can cause problems with roundoff error, so these are dropped.

Table 1. Summary of steps in the variability generation algorithm described in section 2

1. Select and fit the mean response model.
2. Construct residual field $\underline{\mathbf{R}}$ by subtracting mean response from ESM output (equation (3)).
3. Orthogonalize residuals with respect to EOF-0 (equation (7)).
4. Perform the EOF analysis on the residual field.
5. Compute the DFT Φ of the residual field's projection coefficients onto the EOF basis.
6. 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 $[0, 2\pi)$.
7. Compute the projection coefficients $\phi^\star$ of the variability field as the inverse DFT of $\Phi^\star$.
8. Compute the variability field as $\boldsymbol{x}(t) = \sum_{i=0}^{N} \phi_i^\star(t)\hat{\boldsymbol{x}}_i$.

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, Analysis, and ~~Discussion~~Validation

3.1 Model output and performance

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 on a midrange workstation. Each temperature field took 3–4 seconds to generate.

Figure 2 shows a single time slice for each of the variability fields (i.e., the temperature field, with the mean response field subtracted out). The time series these slices were taken from could be used as an ensemble to study the effects of variability on the downstream models that are consumers of these sorts of climate projections.

The spatial structure in the variability is readily apparent. Temperature perturbations occur on scales of roughly 40–60 degrees of arc. Some features, such as the one seen in the low-latitude eastern Pacific, appear in all of the frames, with greater

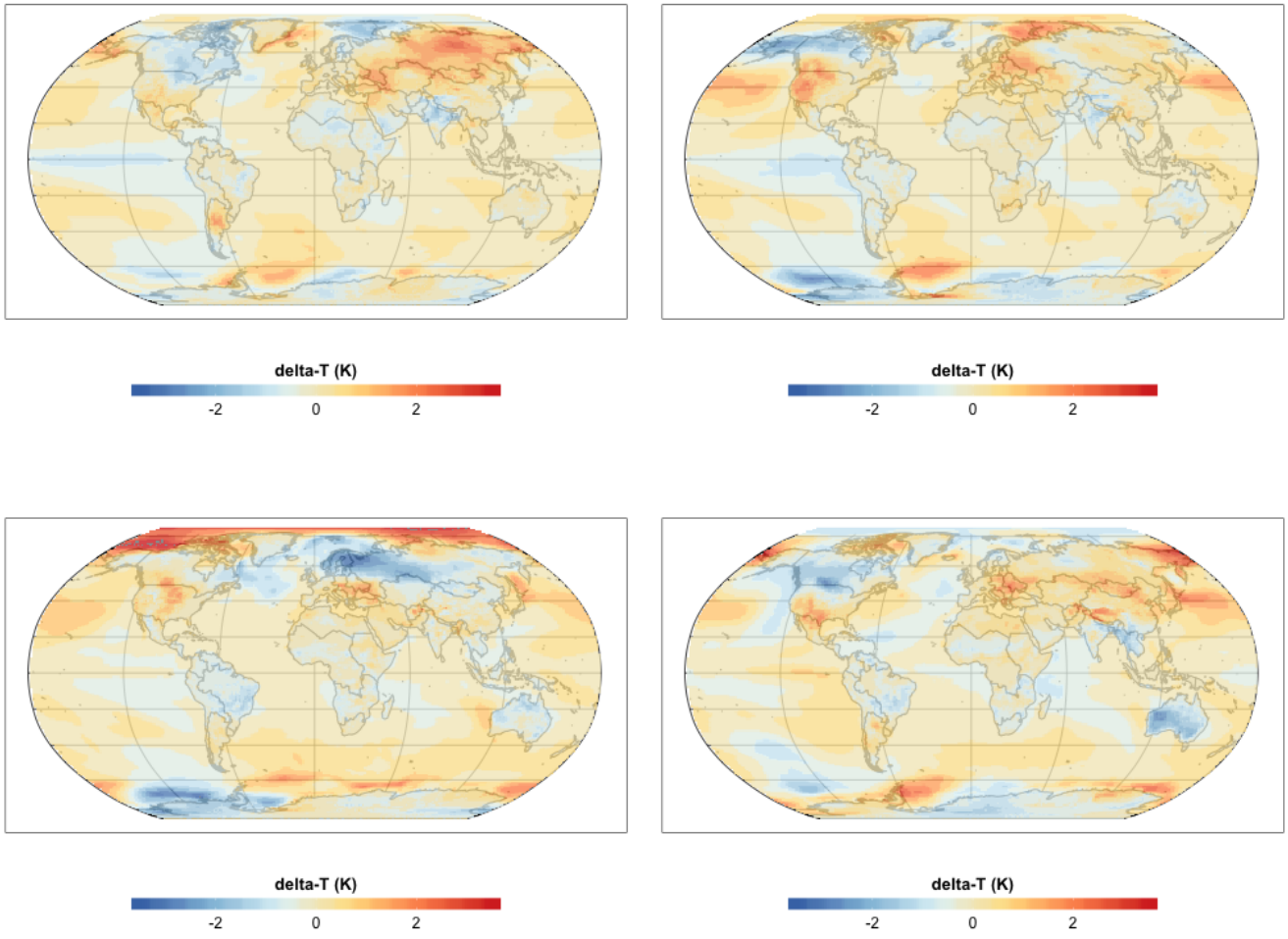

Figure 2. Year 2025 snapshot for variability fields generated using the procedure described in section 2.4. Each field is a different randomly generated realization of the temperature field's departure from the mean response (sec. 2.3). The sequences these frames were drawn from could be used as an ensemble of future climate scenarios for studying sensitivities or uncertainties in models that use climate data as inputs.

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. 3). Two trends are immediately apparent. First, the total power present in each EOF decreases dramatically after the first few EOFs (fig. 3). 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 makes a nontrivial contribution to the result. The last 400

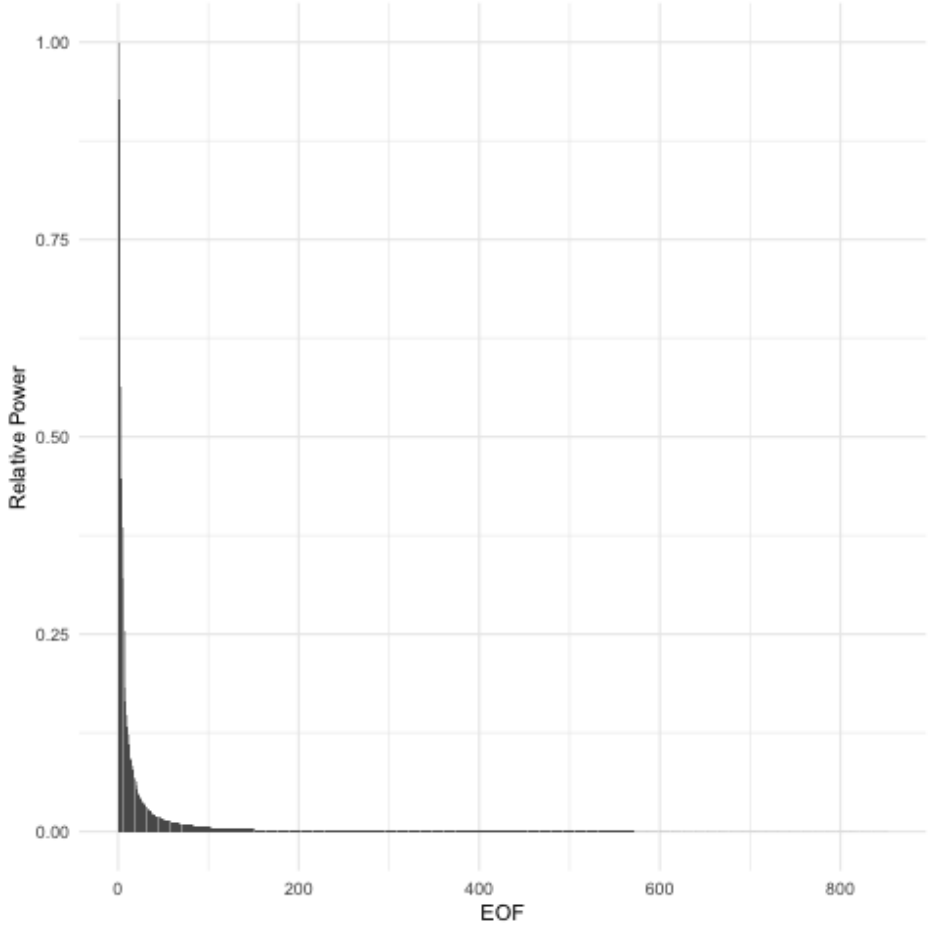

Figure 3. Relative power for each EOF. Roughly half of the total power is contained in the first 10 EOFs. The aggregate power for all EOFs beyond 400 is 1% of the total.

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. 4), 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.

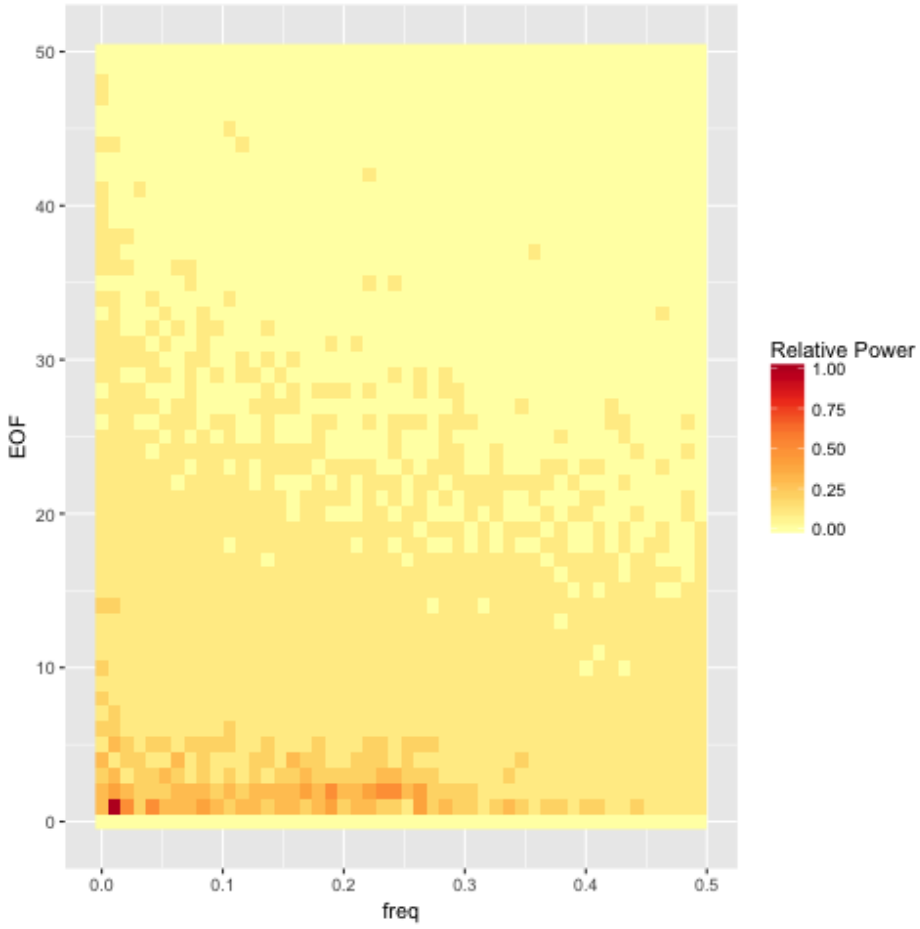

Figure 4. Heat map of power spectral density (PSD) for the first 50 EOFs. The trend of decreasing total power and more uniform spectral density continues for the remaining EOFs beyond EOF-50.

In Figure 5 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~~ 5 years.

Figure 6 visualizes the spatial patterns represented by the first 6 EOFs, and Figure 7 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.

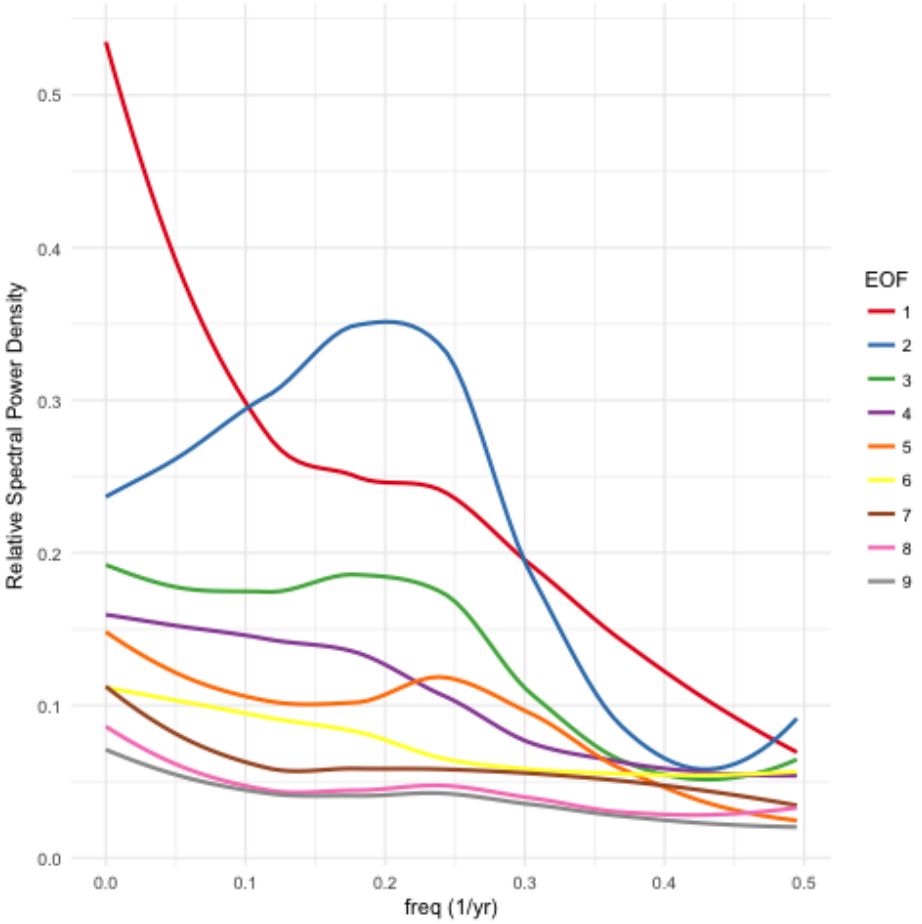

Figure 5. ~~Power~~ Smoothed power spectral density (PSD) for the first 9 EOF basis functions. EOFs 2, 3, and 5 show peaks in the PSD, indicating quasiperiodic behavior on 3–5 year time scales. EOF-1 has most of its power at low frequencies, indicating that this component is approximately (though not exactly) constant over the course of a single ESM run.

3.2 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:

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.

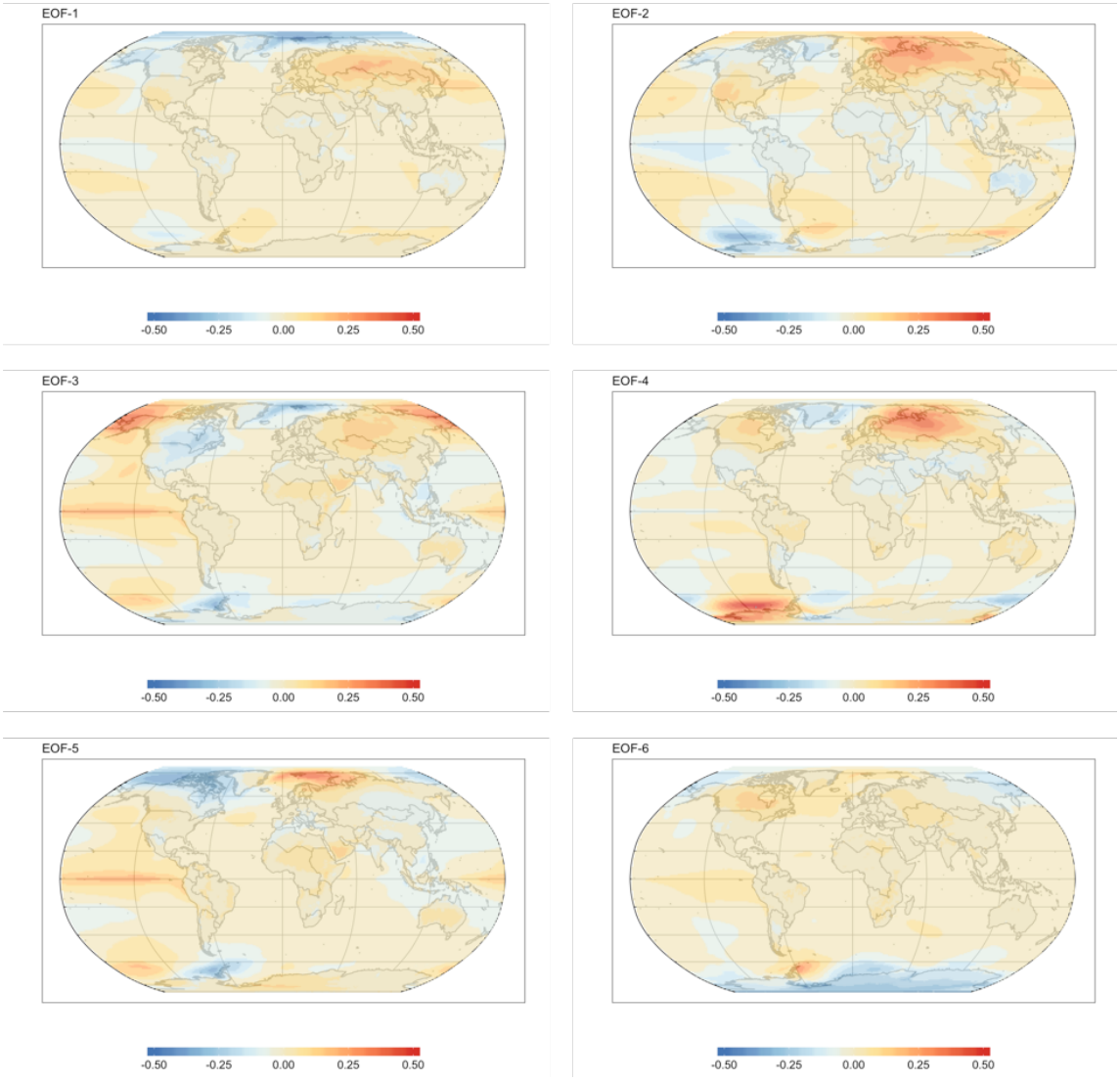

Figure 6. Spatial visualizations of the EOF1-6 basis functions. EOF grid cell values are scaled such that the magnitude of the largest value is 1. These components capture large-scale patterns of variability. EOFs 2, 3, and 5 all feature a temperature anomaly in the eastern Pacific. These same components can be seen in figure 5 to have some periodicity on 3–5 year time scales, suggesting that they may be rooted in physical processes in the ESM the model was trained on.

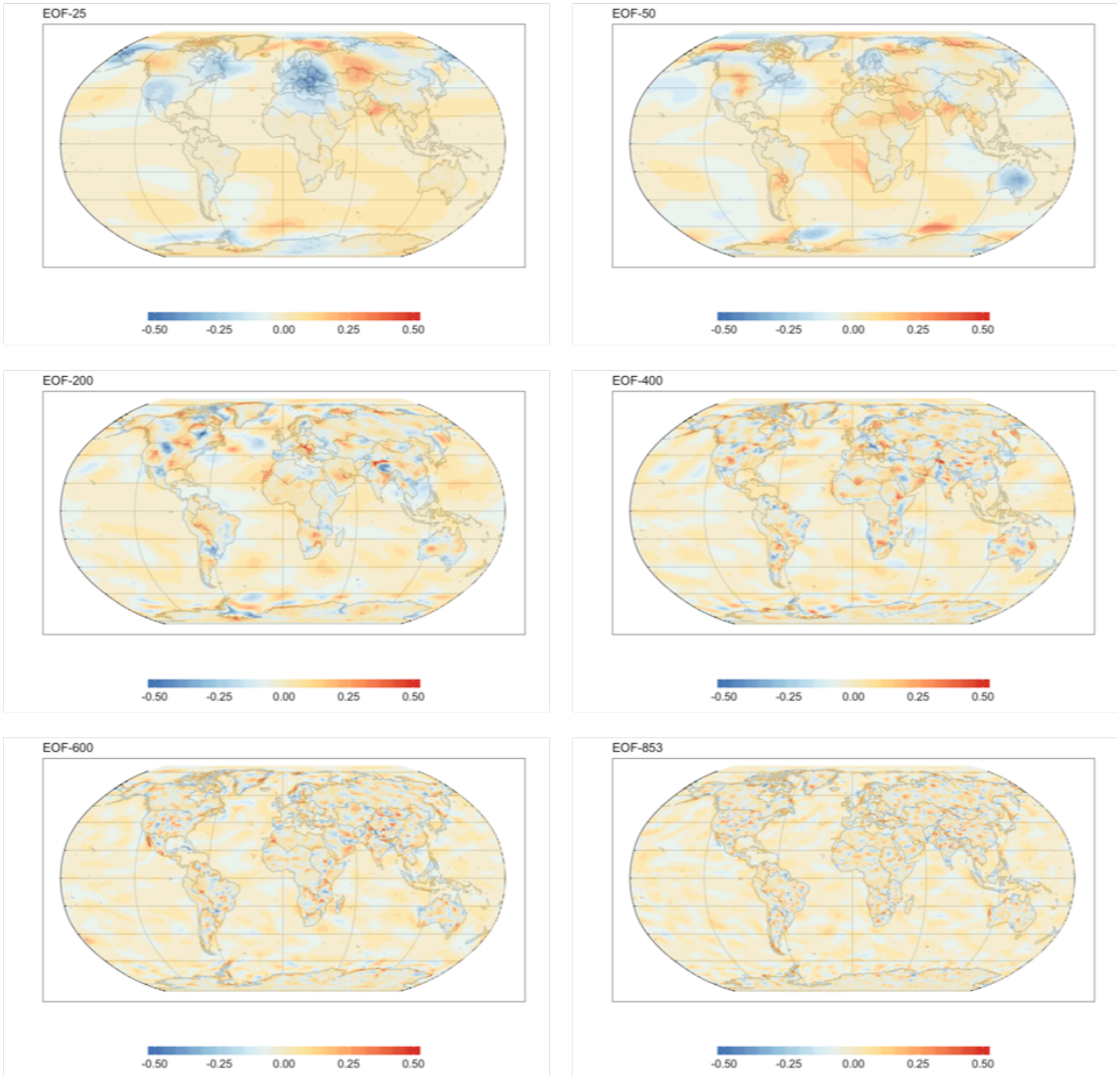

Figure 7. Spatial visualizations of higher EOF basis functions. EOF grid cell values are scaled such that the magnitude of the largest value is 1. The characteristic scale of temperature fluctuations decreases for functions later in the series. Thus, EOFs 25 and 50 show features at about half the scale of those shown in figure 6, while features in EOFs 200 and 400 are roughly one quarter the scale. By the time we get to the last few hundred EOFs, features are just a few grid cells in size, resulting in patterns that might be thought of as spatially structured noise.

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

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.

3.2.1 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.

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.

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 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.

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

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 model 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 $\hat{\boldsymbol{x}}_i$. 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:

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

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

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 $\phi_i^\star$),

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

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

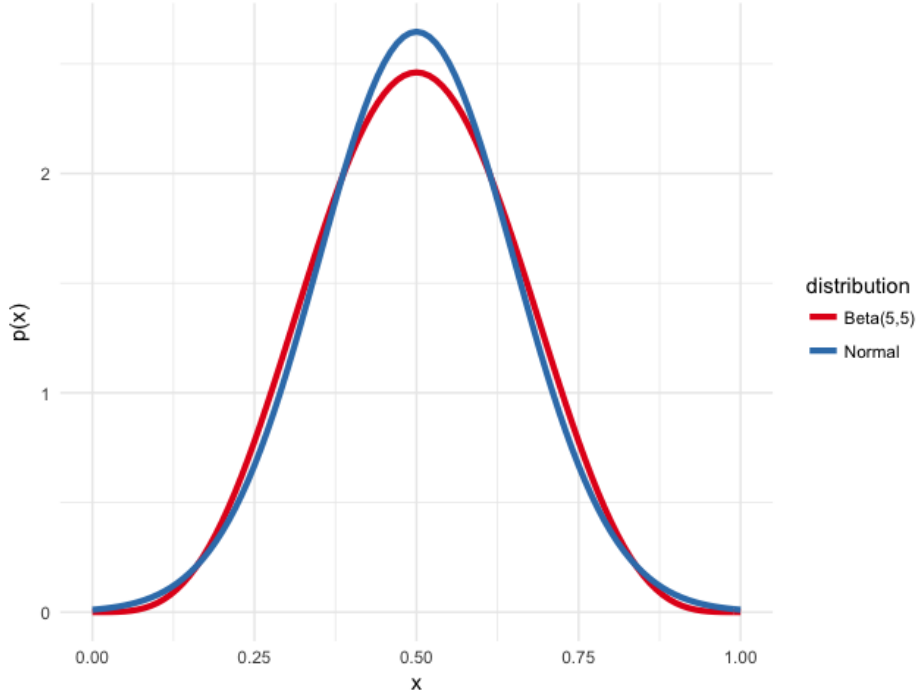

Figure 8. Comparison of the Beta(5,5) distribution and a Normal distribution with equal variance. The beta distribution is zero outside of the depicted range, while the normal distribution asymptotically approaches zero. Although the difference between these two distributions is small, the Shapiro-Wilk test can easily distinguish them.

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 8, 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.

3.2.2 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 3.1 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. 4) was unexpected and might be of interest for further study.

3.3 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 5. 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.

3.4 Underfitting the mean response

Several readers of early versions of this work questioned the decision to fit the mean response model over the entire range of RCP scenarios, speculating that this practice would result in a mean response model that represented a sort of compromise amongst the various RCPs in the input data, fitting none of them particularly well. If the mean response model were to be underfit in this way, then the residuals from the misfitting would be lumped in with the variability and subjected to the randomization procedure described in section 2.4. It was suggested that the long-period behavior of EOF-1 might be evidence that this was happening.

Throughout the rest of this section we will refer to this collection of hypotheses as the *Compromise Conjecture*, or CC for short. We know that the CC is true to some extent, since it seems unlikely that the relationship between global and local temperatures in these models has *no* dependence on the specifics of the warming scenario. One solution to the CC would be to fit separate emulators for each of the RCP warming scenarios; however, for scenarios that do not correspond exactly to an RCP, we still need to generate fields using an approximate mean response, and we will need to know how much of an error we are making. Therefore, the question we must answer is, are the effects of using a compromise model acceptably small in the context of the other approximations used in the emulator's design?

To investigate this question, we fit two more emulators to subsets of the data. The first of these used only the three ensemble members for the RCP-8.5 scenarios. We designated this emulator "RCP85". The second fit used three ESM runs covering the RCP-2.6, RCP-6.0, and RCP-8.5 scenarios. We designated this emulator "MULTI". Our first test was to compare the mean field models for these two emulators. Figure 9 shows a grid cell by grid cell comparison of the w (linear) and b (intercept) coefficients for the two models, from which it can be see that the two mean response models are very similar.

We can quantify just how similar the two models are by fitting linear models predicting the RCP85 coefficients from the corresponding MULTI coefficients. When we do this, we find that the average ratio between the RCP85 and MULTI w terms is 0.994, with an R^2 of 0.999. Most of the residuals are within +/- 0.02 of 0 (for a coefficient that ranges approximately from

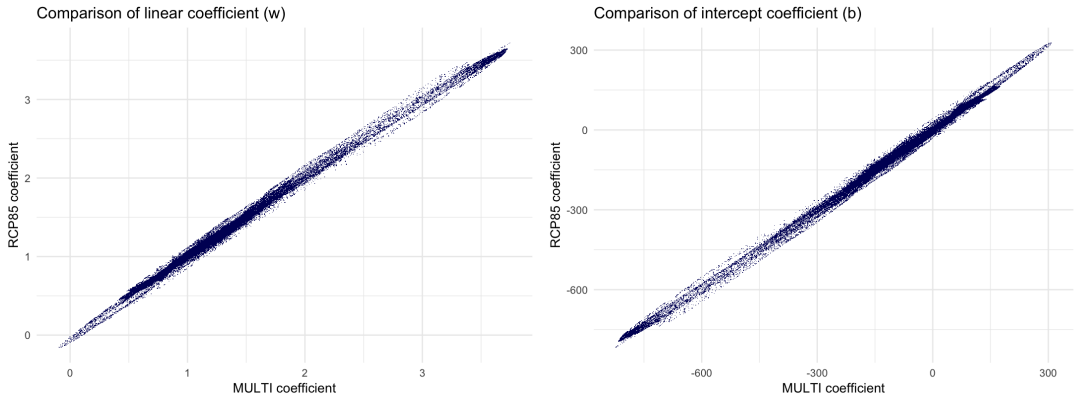

Figure 9. Comparison between coefficients of the mean response model for the RCP85 and MULTI emulators. For both the linear term w (left) and the intercept term b (right) the two models are nearly identical.

0–3). For b, the relationship is nearly as good; the coefficient ratio is 0.987, with an R^2 of 0.998. Most of the residuals are between -5 and +6 (the scale of this variable is considerably larger than the scale for w: approximately -650 – $+300$.)

From this result alone, we see that the mean response models for these two emulators are virtually identical, making it extremely unlikely that CC effects are an appreciable source of error in the MULTI emulator. For this reason, description of additonal tests of CC effects, along with source code and results have been relegated to the data and analysis code archive cited in section 4.

3.5 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 7 and 8 in Table 1.

4 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 and installation instructions 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.1183640.

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~~ as part of research in the Multi-Sector Dynamics, Earth and Environmental System Modeling Program. The Pacific Northwest National Laboratory is operated for DOE by Battelle Memorial Institute under contract DE-AC05-76RL01830.

This research was supported in part by the Indiana University Environmental Resilience Institute and the Prepared for Environmental Change grand challenge initiative.

References

Akhtar, M. K., Wibe, J., Simonovic, S. P., and MacGee, J.: Integrated assessment model of society-biosphere-climate-economy-energy system, Environmental Modelling & Software, 49, 1 – 21, https://doi.org/https://doi.org/10.1016/j.envsoft.2013.07.006, http://www.sciencedirect.com/science/article/pii/S1364815213001655, 2013.

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.

Calvin, K. and Bond-Lamberty, B.: Integrated human-earth system modeling—state of the science and future directions, Environmental Research Letters, 13, 063 006, 2018.

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.

Cui, Y., Calvin, K. V., Clarke, L., Hejazi, M., Kim, S., Kyle, G. P., Patel, P., Turner, S. W., and Wise, M.: Regional responses to future, demand-driven water scarcity, Environmental Research Letters, 2018.

Elliott, J., Deryng, D., Müller, C., Frieler, K., Konzmann, M., Gerten, D., Glotter, M., Flörke, M., Wada, Y., Best, N., et al.: Constraints and potentials of future irrigation water availability on agricultural production under climate change, Proceedings of the National Academy of Sciences, 111, 3239–3244, 2014.

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.

Nelson, G. C., Valin, H., Sands, R. D., Havlík, P., Ahammad, H., Deryng, D., Elliott, J., Fujimori, S., Hasegawa, T., Heyhoe, E., et al.: Climate change effects on agriculture: Economic responses to biophysical shocks, Proceedings of the National Academy of Sciences, 111, 3274–3279, 2014.

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.

Ray, D. K., Gerber, J. S., MacDonald, G. K., and West, P. C.: Climate variation explains a third of global crop yield variability, Nature communications, 6, 5989, 2015.

Rosenzweig, C., Elliott, J., Deryng, D., Ruane, A. C., Müller, C., Arneth, A., Boote, K. J., Folberth, C., Glotter, M., Khabarov, N., et al.: Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison, Proceedings of the National Academy of Sciences, 111, 3268–3273, 2014.

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.

Turner, S. W., Hejazi, M., Kim, S. H., Clarke, L., and Edmonds, J.: Climate impacts on hydropower and consequences for global electricity supply investment needs, Energy, 141, 2081–2090, 2017.

Voisin, N., Hejazi, M. I., Leung, L. R., Liu, L., Huang, M., Li, H.-Y., and Tesfa, T.: Effects of spatially distributed sectoral water management on the redistribution of water resources in an integrated water model, Water Resources Research, 53, 4253–4270, 2017.