Fldgen v1.0: A Computationally Efficient ~~Emulators~~ Emulator 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)

Authors' responses to reviewers' comments

Anonymous Reviewer #1

Major comments

1. Please don't use 'emulator' at all in your manuscript. What you have implemented is a 'surrogate model' or 'metamodel'. An emulator is a type of surrogate model / metamodel which is an interpolator and gives a probability distribution for outputs corresponding to inputs it is not trained at. See [1] for details. I understand that others in the earth system modelling community have used 'emulator' in the same way as you, but whoever started using it first and passed down this definition is incorrect for doing so.

The reviewer's point is well taken; however, as the reviewer points out, the term "emulator" is firmly entrenched in the earth system modeling community. Moreover, it is the term used for this kind of model in other papers in this journal. Therefore, we have elected to stick with the terminology that is customary among our target audience.

2. Your literature review in the introduction is very limited. This would be fine if the journal was very narrowly focused, but GMD has a very broad appeal. Statisticians from the UQ (Uncertainty Quantification) community have, for example, done a lot research on surrogate modelling methods for very expensive models like an ESM. Names that come to mind are Nathan Urban (Los Alamos), Jonty Rougier (Univ. of Bristol), Michael Goldstein (Durham univ.). A series of workshops were held in Cambridge earlier this year (http://www.newton.ac.uk/event/unq/workshops) which will help you find statisticians working in this field. Non-statisticians like David Sexton at the Met office in the UK are also working on quantifying uncertainty of ESMs. There are probably lots of other research that you can also mention, but above is just a start.

When preparing the final draft of our manuscript, we will broaden the discussion of related literature.

3. Please use more standard mathematical notation for defining vectors and matrices. In particular, please don't use $|x\rangle$ and $\langle x|$ for column and row vectors. I have never seen this notation being used before. It is much more standard to use x for a column vector and xT (i.e. the transpose of x) for a row vector.

In the final draft we will replace the Dirac notation currently used in the manuscript with $\boldsymbol{x}$ and $\boldsymbol{x}^T$.

4. Please also stick to normal conventions for matrix and vector algebra, e.g. page 2 line 28 you state "Occasionally we will add a matrix and a vector e.g. $B = A + |x\rangle$. Please do not do this! I know that you explain what you mean when you do this, but mathematically it is not the correct way of doing things because you're effectively defining $|x\rangle$ to be a vector sometimes and a matrix and other times. If you want to add a matrix with a vector in the way you describe then define a new matrix which has as its columns or rows the vector you want. This is a much clearer way of defining things.

In all cases $\boldsymbol{x}$ (formerly $|x\rangle$) is a vector. Writing $\mathbf{B} = \mathbf{A} + \boldsymbol{x}$ is merely a shorthand way of saying $\mathbf{B} = \mathbf{A} + \mathbf{M}(\boldsymbol{x})$, where $\mathbf{M}(\boldsymbol{x})$ is the outer product of $\boldsymbol{x}$ with a suitably-sized vector of ones (i.e., $\mathbf{M}(\boldsymbol{x}) = \mathbf{1}\boldsymbol{x}^T$). Making this operation explicit doesn't make the discussion any clearer; quite the contrary, it forces the reader to stop and figure out what the new operator actually does. Likewise, defining a new matrix each time we need to perform this operation proliferates symbols unnecessarily and obscures the relationship between the vector and matrix versions of the same quantity. For these reasons, we have elected to keep the broadcast notation.

5. Page 3, line 8. You state "We refer to the ESM data as "observations" . . . ". NO, NO, NO! Please do not do this. I spent about an hour reading your paper thinking you were using real observations and then I realized I had not properly read this very important line in your methods. If you want to use your ESM output as measurements then call them "synthetic measurements". Everyone knows what this is. But if you just state measurements or observations, we think you're using the real thing.

In the final draft we will refer to the ESM output as "synthetic measurements".

6. There are lots of subject specific words / jargon used here which you assume that the reader knows the meaning of because you don't define them. Examples include: 'linear pattern scaling', 'discrete Fourier Transform', 'randomizing the phases of G', 'spatial coherence', etc . . .

In the final draft we will include explanations of any terminology that might be unfamiliar to a broader scientific audience.

7. For things that you define you don't give enough detail. For example, with EOFs (statisticians use principal components but it means the same thing) you need to say that we normally only choose the first n EOFs where n is determined such that most of the variance (or power as you mentioned) is explained. Only when we get to the results section do you talk about the number of EOFs that you're using.

We thank the reviewer for pointing out this important omission. Although principal components are often used for dimensionality reduction, in this case we are using them solely to diagonalize the covariance matrix. The final draft will include an explanation of this difference.

8. I was also completely lost with section 2.4. You seem This seems to be key part of the methods, so really needs to be explained better. You make statements that make no sense to the non-specialist – e.g. in the lines prior to equation 7 you state that you're making some minor modifications to the standard procedure, but why? And what is a zeroth basis vector and why are you defining it in this way. These things I read and go "okay", but I have no idea why you're doing this. This is just one example of many scattered throughout the manuscript.

The purpose of including it is explained in detail in the paragraph following equation 7. ("This property guarantees that .... This property is useful because ....") We will move this paragraph up before equation 7 so that the differences between this basis vector and the others and the motivation for including it at all are clear before we begin describing how it is calculated.

9. There's no motivation or justification for doing what you're doing. At the end of the introduction, you state that there's a need to quantify the uncertainty in ESM output. Fine. But then you jump straight into your approach of quantifying this uncertainty by generating it based on an error covariance matrix that is derived from defining the temporal and spatial correlation of the ESM output. Why is this a good way of defining the ESM uncertainty?

Our paper is not about quantifying uncertainty in ESM output at all. Instead, it is about providing a source of data, beyond what is available in public archives, for models that are *consumers* of ESM output. Uncertainty studies (in these models, not in the ESMs themselves) are just one reason why we might want to generate these datasets (we give two others in the introduction). In the final draft the opening paragraphs of the introduction will be rewritten to clarify these points.

10. Following on from the previous point, this seems to be a major flaw in this paper. Normally when we do uncertainty analysis, we propagate the uncertainty from the inputs (e.g. uncertainty in model parameters) through to the uncertainty in the ESM outputs. Instead, you seem to be using the spatial and temporal correlations of the ESM output as a means to generate the uncertainty in the ESM output with the metamodel (or "emulator" as you call it) as the vehicle for carrying out the extra ESM runs. Perhaps I'm mistaken and this isn't what you're doing. If I am mistaken then the fact that I have misunderstood this is a major problem. If you want people to be interested in your research, you first need to communicate it clearly and (sometimes) simply to them.

As we noted above, the purpose of this paper is not to study uncertainty in ESM outputs. It is important to realize here that the variability in ESM outputs is an important contributor to the uncertainty in downstream models *and would still be so even if there were no parameter uncertainty at all in ESMs*. Indeed, for many downstream models the uncertainty contribution from variability is *more* important than that from parameter uncertainty. As we discuss in the introduction, many ESM emulators do not produce this variability, or produce it with significant limitations. The purpose of this work is to produce an emulator that produces the full range of variability seen in the ESM data and does so in a realistic way.

The revised manuscript will explain these matters in more detail than the current manuscript.

11. I am confused with how you train your metamodel (or "emulator" as you call it). When I train metamodel, this often involves multiple runs of the computationally expensive model (i.e. the ESM in your case). In your paper, it seems that you just need one run of the ESM to train the metamodel. Is this correct? Again, this isn't how we normally train metamodels so you need to be really clear about this. In fact there should really be a whole section in the methods explaining everything about how you constructed the metamodel. Maybe you feel what you've written is enough, but I'm just generally confused so you need to lay things out much more logically and clearly at the very least.

In section 2.2 (p.3, l. 5–7) we wrote:

> 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.

12. Table 1 makes little sense to me. If I really concentrate I can probably understand what's going on, but you could help the reader by using less complicated words or phrasing it in a simpler way.

Table 1 is a summary of the steps in the algorithm described in the rest of the section. Although the reviewer appears not to have found it useful, other people we have shown it to have said that the summary helped them visualize how the pieces of the algorithm fit together. Therefore, we are inclined to keep it. In the final draft of the manuscript we will expand the table caption to clarify that this is a summary of the material in the rest of the section.

13. The results section seems to be too short (less than one page). Most papers I read have at least 3 or 4 subsections within the results section. These subsections have their own titles and help navigate the reader through the different aspects of the results. At the moment, the results just seems like a list of things. Your results show [sic] flow more like a story. You also don't really give a lot of detail, e.g. just one sentence for figure 5? What's the point of having it in there?

In dividing the paper into sections, we construed "Results" narrowly to mean "the artifacts produced by running the model". Conversely, we categorized analysis of the properties of the model output as "Discussion", which is why that section comprises 10 of the paper's 17 pages, with three subsections and two sub-subsections. In the final draft we will merge the two sections into a single "Results and Discussion" section.

In addition, the final draft will describe some of the salient features in the map figures.

14. I didn't really read through the discussion in depth, but in section 4.1.1 (and may other subsections) you describe further results that were carried out. Thse should be in the results section. The purpose of the discussion section is: (a) to give an explanation for why your results look the way they do; (b) to put your results in context of other comparable studies. I see a bit of (a) in the discussion but no evidence of (b).

As explained above, we appear to have a different convention than the reviewer regarding the distinction between "results" and "discussion". In the final draft the two will be merged into a single section.

Minor comments

- Panels of figures: label them with letters. E.g. Figure 1a would refer to panel at the top left of figure 1.

These will be added in the final draft.

- Figures: captions lack enough details

The captions will be expanded in the final draft.

- When submitting for review, it's more helpful to put all the figures and tables at the end of the manuscript. This makes it easier for the reader to refer to a particular figure when reading a particular part of the results.

Opinions differ on this; many scientists prefer in-text figures and tables. The GMD author guidelines leave it up to the author.

Response to Anonymous Reviewer #2

General comments

- However, the language used is quite mathematical for a GMD paper. I think this could be addressed without loss of quality or conciseness.

When preparing the final draft of the manuscript, we will look for opportunities to reduce the density of the mathematics in the text. We do note, however, that it was important to us to provide enough detail for readers to be able to both recreate and evaluate the algorithm for themselves, if desired, and doing so requires a certain amount of mathematical specificity.

- Also, as suggested by the first reviewer, this is not an emulator in the strict sense.

It would seem that there is some diversity in the way this terminology is used in different scientific communities. Amongst the researchers who develop these kinds of models the term "emulator" seems to be preferred; therefore, we have elected to keep to that convention.

- I also agree with the first reviewer in that, ESM outputs are not "observations". "ESM outputs" would suffice.

We have adopted the first reviewer's suggestion of "synthetic measurements" to refer to the data being used to train the model.

- A related point is that the model simulates global mean surface temperature from GCMs (general circulation models/global climate models - choose your favourite acronym) rather than ESMs. The CMIP5 definition

of an ESM includes an interactive carbon cycle, going from emissions to concentrations to forcing to temperature. GCMs skip the emissions step, running from prescribed concentrations that have been calculated from a simple model, e.g. MAGICC, as they were in CMIP5.

There is nothing in the model that is specific to GCMs as contrasted with ESMs. The particular input data we chose to use as a demonstration were forced by concentration, but we could equally well have selected archival datasets that were produced with the carbon cycle turned on. Since the developers of CESM refer to their model as an "earth system model", we chose to do the same, even when working with scenarios run in a mode more characteristic of a GCM.

Specific comments

- In the introduction, the application of the model to extreme events is given as a justification for its creation. However, the model only produces annual mean temperature output in each grid cell. I am not aware of an extreme indicator that uses annual mean temperatures. Such indicators are usually calculated from daily climate model output (see Zhang et al 2011, 10.1002/wcc.147). This would be a natural extension to this model, but in its current form it is not capable of analysing "extremes" in the usual sense.

This is a good point. By "extreme events" we had in mind the tails of the distribution of annually averaged values. We will adjust the language in the final draft to clarify what we had in mind.

- I don't disagree with the authors about the notation convention: I understand the broadcasting concept used in their convention and agree it aids readability. I do find it hard to follow the equations though. If we have $|T_g\rangle = \mathbf{O}\,|\lambda\rangle$, then this suggests to me that $|T_g\rangle$ is a column vector of shape 855 x 1 formed by multiplication of **O** (855 x 55296) by $|\lambda\rangle$ (55296 x 1). In eq(2) you have $T_g\,|w\rangle + |b\rangle$. Is T_g (not bracketed in eq(2)) times $|w\rangle$ a column vector times a column vector? ? How is this defined?

T_g (without brackets) is a scalar. On the other hand, $|T_g\rangle$ is a vector of global mean temperature values. When defined by the first equation in the quote, this vector is made up of the values of T_g for each year of each model in the input set. In other words, the name of a variable tells us what physical quantity the variable represents, and the decoration tells us how many we have and what kind of structure they are organized into. We will add some clarifying remarks on this point to the notation section.

- And then in equation 3, there is $|T_g\rangle$ (a column) times $\langle w|$ (a row), which I think is 855 x 855, then added to $|b\rangle$ (855 x 1)? and subtracted from **O** (55296 x 855 - but how is this broadcasted?) If there are no typos in these equations, it would be helpful here to put in a diagram of the matrix dimensions in the equations 1 to 3.

The symbol $\langle w|$ is a row vector, with dimension 1 x 55296 (i.e., one value for each grid cell). The product $|T_g\rangle\,\langle w|$ is an outer product, the result of which is a matrix $(855 \times 1) \cdot (1 \times 55296) = (855 \times 55296)$. The vector $|b\rangle$ likewise has dimension (55296 x 1) (again, one value for each grid cell). Because this matches the number of columns in the matrix formed by the outer product, it can be broadcast in the usual way. The result is still a matrix (855 x 55296), which is confomant with the matrix **O** that it is being subtracted from.

We will clarify the dimensions of the vectors of pattern scaling coefficients, and we will add a figure that shows how these quantities fit together to produce the final matrix of residuals.

– σ values in table 1 and p5 line 9. I think these are the singular values of $\mathbf{R}$, but it is not really explained what these are or what they mean. This paragraph could do with some expansion of the key terms (rank deficient, discrete Fourier transform). Does dropping EOFs where $\sigma < \sigma_{\text{threshold}}$ guarantee full rank?

The σ are the singular values; we will clarify this in the final draft. We will also provide a brief explanation of what the singular values mean, and we will supply a reference to an approachable introduction to Fourier transforms and their applications.

Technically, having all $\sigma > 0$ is enough to guarantee full rank, so it would be more correct to say that the problem here is ill-conditioning, rather than rank deficiency. However, because we do not use the SVD to invert the matrix (only to find the principal components), it is not clear that the ill-conditioning causes any particular harm. Therefore, in the final draft we will regard the dropping of components with very small singular values as an implementation detail and omit the discussion of it in the text.

– Section 3: Can the four images in figure 1 be interpreted as ensemble members? If so, it would be good to state this.

Yes, they can. We will comment on this in the final draft.

– figures 4-6 and associated discussion in lines 24-28 on page 6: The periodic variability in EOFs 2, 3 and 5 - could these have a physical interpretation? For example there seems to be an El Nino style feature in EOFs 3 and 5. On the other hand, is there any evidence that the lower EOFs are not just noise?

We, too, had noticed the resemblance to El Nino in those components; however, it wasn't clear how to make a rigorous comparison between the patterns we see here and real-world El Nino events (since surface air temperature isn't really the right variable for computing a proper El Nino index). Developing a methodology for making such a comparison is outside the scope of this paper (though it would be interesting research in its own right), so we decided to characterize these components generically as periodic modes of variability, rather than to attribute a physical cause to them.

We feel very confident that the lower EOFs (we assume that by this you mean the ones with lower total power, not the ones earlier in the sequence) *are* mostly noise. In the time dimension their power spectra are almost completely flat, and the length scale of spatial correlations is just a few pixels. This is pretty much the definition of "noise" in this context. That said, there is still *some* structure, even in these noisy basis functions, and characterizing that structure with this model allows us to ensure that the noise in the output realizations has the *same* structure.

To put it another way, you could probably get an adequate representation of the noise in the system by just applying a random perturbation (i.e., without regard to space or time correlation) and then running a smoothing kernel over the result so as to reproduce the short-range correlations observed in the noisy components. But, what should be the width

of that kernel, and how should that noise field be weighted relative to the structured components? Those things are an important part what we are trying to model with this technique.

- 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.

We have worked with models trained on a single scenario, and for the most part the results are qualitatively similar to the multi-scenario results. We didn't try to run any statistical tests to detect differences, but with the limited amount of data available it seems unlikely that any such differences would be detectable. Therefore, although it's theoretically possible that by using variability from a model trained solely on a scenario of interest (supposing you know in advance what scenario that is) you might get more accurate results for that scenario. However, in practice the difference is likely to be small and perhaps offset by the effects of having less data to train on. Many of these topics would be worth revisiting in the future, particularly once improvements in the mean field response are in place.

Technical Corrections

- page 5, line 3: allow → allows

Thanks. We will correct this.

- page 6, line 3: 143 seconds. What is the machine architecture here?

This was on a midrange workstation. We will mention this in the final draft.

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~~ 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 ~~uncertainty studies and similar uses~~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 ~~have utility~~ be useful in the study of climate ~~uncertainty and variability , extreme events, and climate change mitigation~~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 ~~need many realizations of one or more~~ use data from 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).~~

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~~; however, running ESMs is difficult and costly. As a result, most ~~researchers~~ 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 ~~tried to produce~~ 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 ~~the small collection of~~ they might need many realizations of one or more future climate scenarios.

Examples of research areas for which archival runs might be insufficient ~~for a robust statistical analysis.~~ 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 in emulators is crucial to understanding the behavior of and uncertainties in these models.

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~~ underlined bold symbols (*e.g.* ~~R~~$\mathbf{R}$) to refer to matrices. ~~Vectors are represented using Dirac notation, $|x\rangle$ for a column vector and $\langle x|$ for a row vector~~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. ~~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.*, ~~$\mathbf{B} = \mathbf{A} + |x\rangle$~~$\mathbf{B} = \mathbf{A} + \boldsymbol{x}$. This should be interpreted to mean that the vector ~~$|x\rangle$~~ $\boldsymbol{x}$ is to be added to each row of the matrix ~~A~~$\mathbf{A}$. Therefore, the length of ~~$|x\rangle$~~ $\boldsymbol{x}$ must be equal to the number of columns in ~~A~~$\mathbf{A}$. 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~~ 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 "~~observations" because they are the data we are trying to emulate. Similarly, when we refer to~~ synthetic measurements" (when referring to the data as a whole) or "cases" (when referring to individual frames in the data), while the terms "results" ~~or~~ and "model output" ~~without further clarification, we are talking about~~ will be reserved for 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~~ synthetic measurements will be grouped into the input matrix ~~$\mathbf{O}$, with observations~~ $\mathbf{O}$, with the cases in rows and grid cells in columns. In the input data used for this study, each ~~observation~~ case is 288 (longitude) $\times$ 192 (latitude), for a total of 55296 grid cells. Therefore, in this case, ~~$\mathbf{O}$~~ $\mathbf{O}$ 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

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

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 ~~$|x\rangle$ is $T_g = \langle\lambda|x\rangle = \langle x|\lambda\rangle$~~ $\boldsymbol{x}$ is $T_g = \boldsymbol{\lambda}^\top\boldsymbol{x} = \boldsymbol{x}^\top\boldsymbol{\lambda}$. Similarly, the matrix-vector multiplication ~~$|T_g\rangle = \mathbf{O}|\lambda\rangle$~~ $\boldsymbol{T_g} = \mathbf{O}\boldsymbol{\lambda}$ 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 ~~observed~~ 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 ~~coefficients $|w\rangle$ and biases $|b\rangle$~~

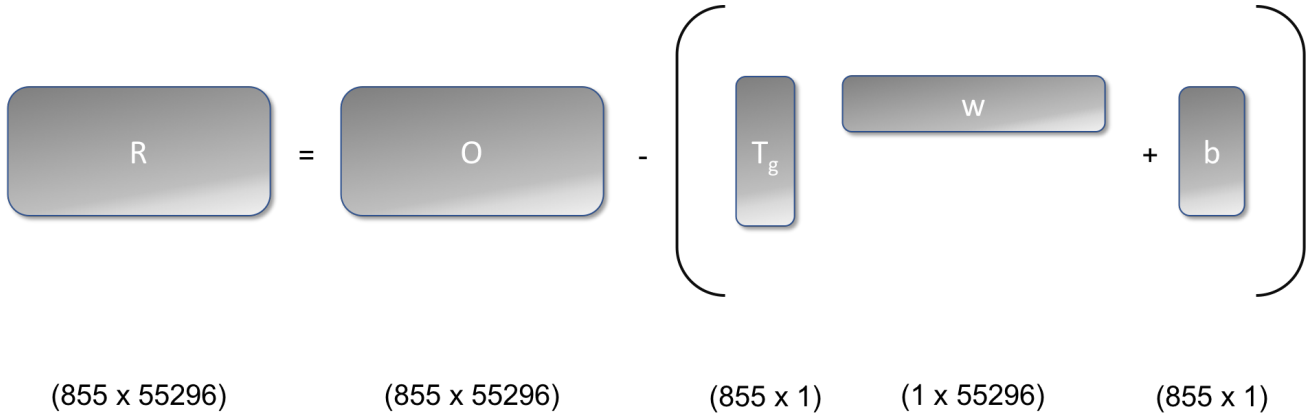

Figure 1. Schematic of the residual calculation showing the shapes of the matrices involved. The result of the outer product $\boldsymbol{T_g w}^\top$ is an 855×55296 matrix. The vector $\boldsymbol{b}$ is added to this matrix using the broadcast convention described in section 2.1

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 ~~(m) for~~ $\boldsymbol{m}$ for global mean temperature T_g is given by

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

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

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

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, ~~R~~$\underline{\underline{\mathbf{R}}}$, 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 ~~magnitude of~~ 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 $|G'| = |G|$, ~~randomizing~~ 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.

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

where

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

The ~~$|\hat{x}\rangle_i$~~ $\hat{\boldsymbol{x}}_i$ 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.

~~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~~ 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 ~~$|\hat{x}\rangle_0$~~ $\hat{\boldsymbol{x}}_0$ to be the global mean operator, normalized to unit magnitude:

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

We force ~~this~~ $\hat{\boldsymbol{x}}_0$ to be a basis vector by subtracting from each residual vector its projection onto ~~$|\hat{x}\rangle_0$~~ $\hat{\boldsymbol{x}}_0$ and performing the SVD on the modified residuals. This procedure ~~guarantees that~~ forces all of the basis vectors ~~have zero area-weighted global mean , or, equivalently,~~ 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 ~~variability in the global meantemperature 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~~ other basis vectors will have zero global mean. Therefore, if $\phi_0(t) = 0$, then the global ~~mean will not be affected at all~~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.

~~In practice **R** will usually be rank deficient or nearly so.~~ 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 ~~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 , $\sigma_{\text{threshold}} = 10^{-8}\sigma_{\text{max}}$~~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.

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 and Discussion

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 ~~snapshot in time~~ single time slice for each of the variability fields ~~. The spatial coherence of the 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 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.

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

1. Select and fit the mean response model.
2. Construct residual field ~~$\mathbf{R}$~~ $\mathbf{R}$ 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. ~~Drop EOFs with singular values $\sigma < 10^{-8}\sigma_{\max}$.~~
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 ~~$|x(t)\rangle = \sum_{i=0}^{N} \phi_i^\star(t)|\hat{x}\rangle_i$~~ $\boldsymbol{x}(t) = \sum_{i=0}^{N} \phi_i^\star(t)\hat{\boldsymbol{x}}_i$.

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 ~~make~~ makes 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. 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.

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 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-

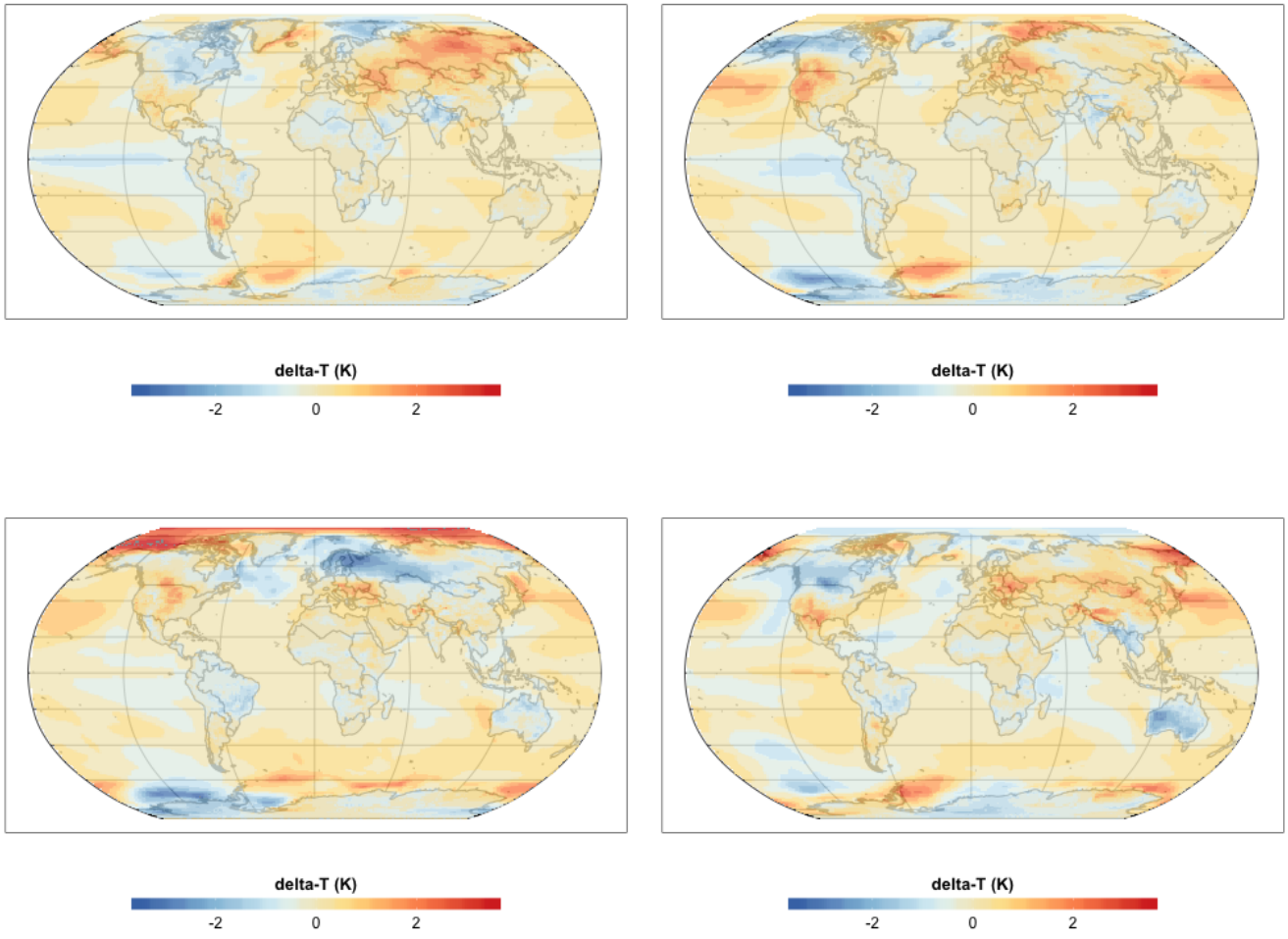

Figure 2. Year 2025 snapshot for variability fields generated using the procedure described in section 2.4. Each field is a different randomly generated realization of the temperature field's departure from the mean response ~~.~~(sec. 2.3). The sequences these frames were drawn from could be used as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

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.

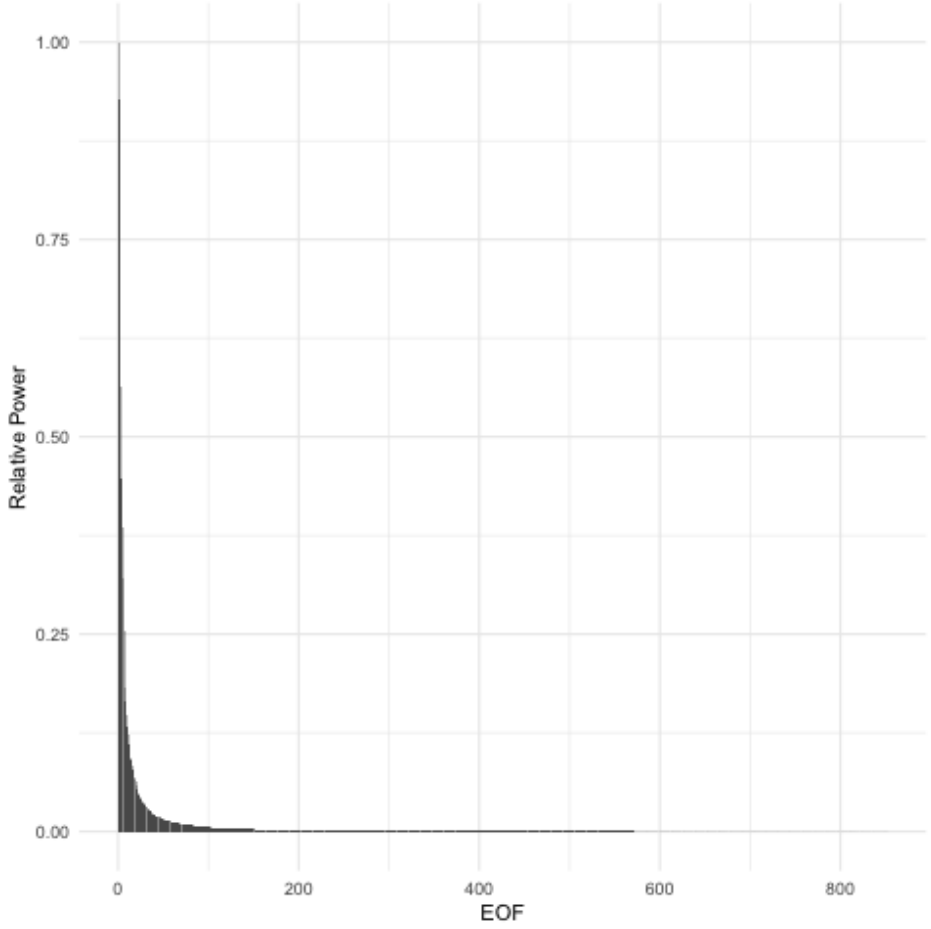

Figure 3. Relative power for each EOF. Roughly half of the total power is contained in the first 10 EOFs. The aggregate power for all EOFs beyond 400 is 1% of the total.

4 ~~Discussion~~

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.

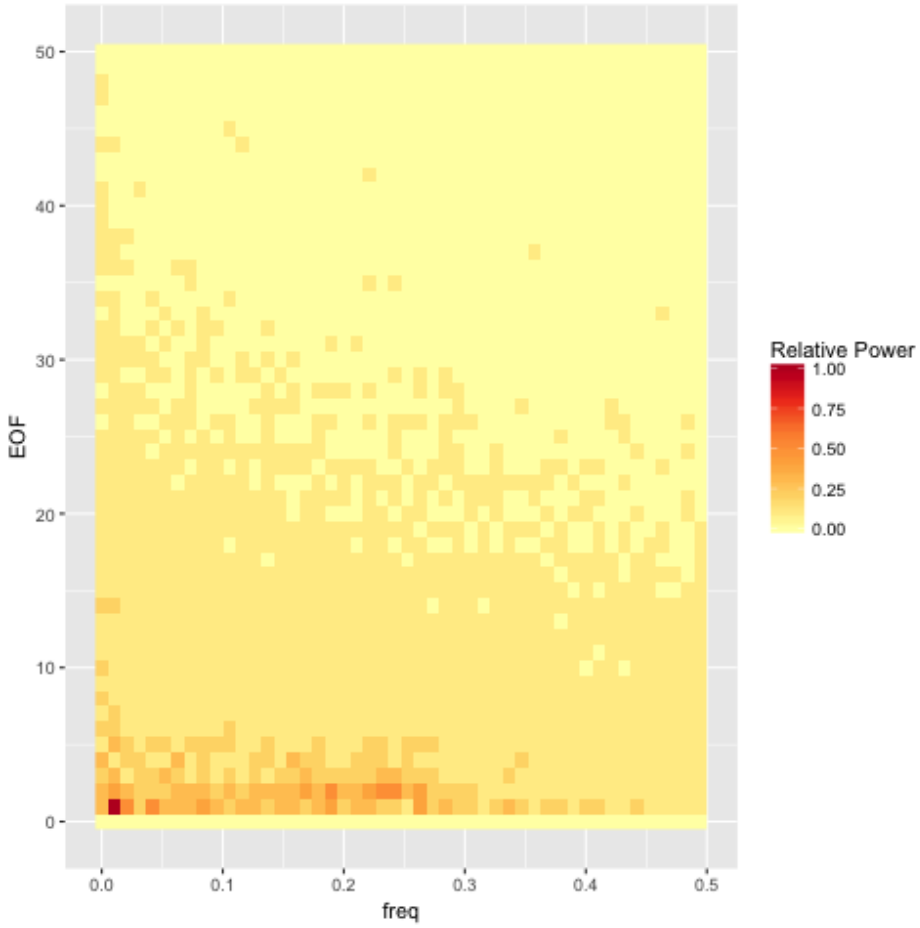

Figure 4. Heat map of power spectral density (PSD) for the first 50 EOFs. The trend of decreasing total power and more uniform spectral density continues for the remaining EOFs beyond EOF-50.

In 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:

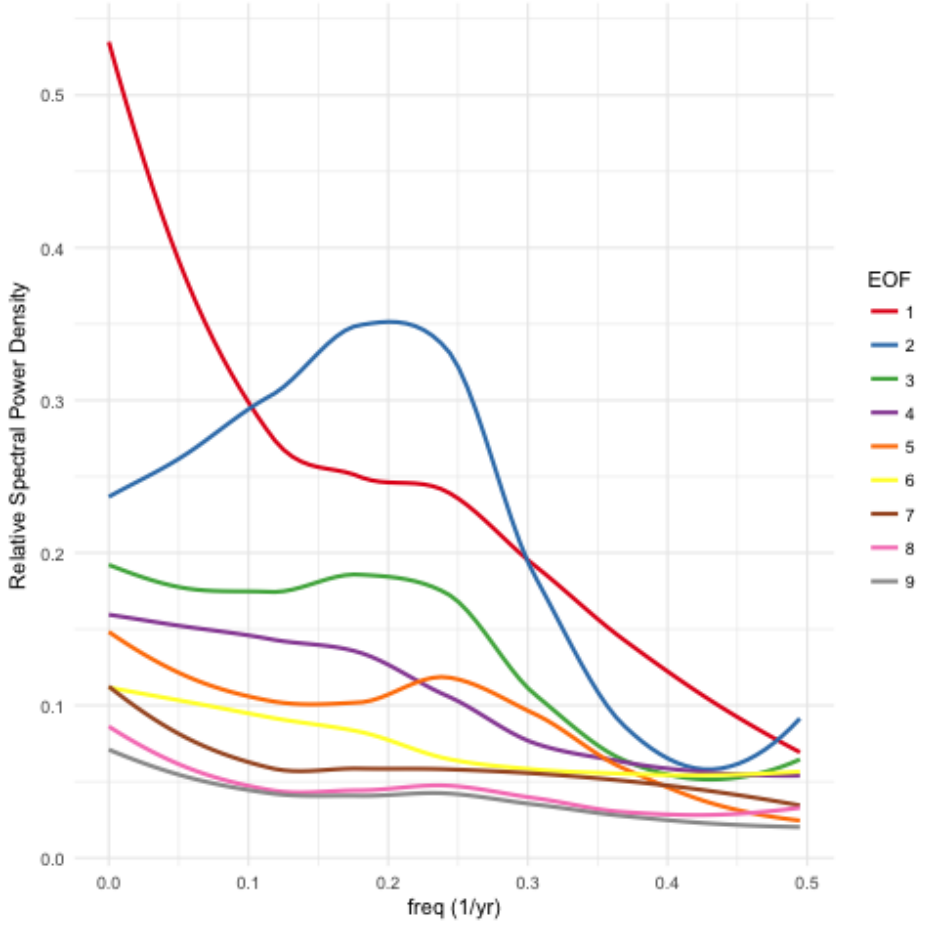

Figure 5. Power spectral density (PSD) for the first 9 EOF basis functions. EOFs 2, 3, and 5 show peaks in the PSD, indicating quasiperiodic behavior on 3–5 year time scales. ~~EOFs 1, 3, 4, and 5 also have~~ EOF-1 has most of its power at low frequencies, indicating that this component is approximately (though not exactly) constant over the course of a single ESM run.

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.

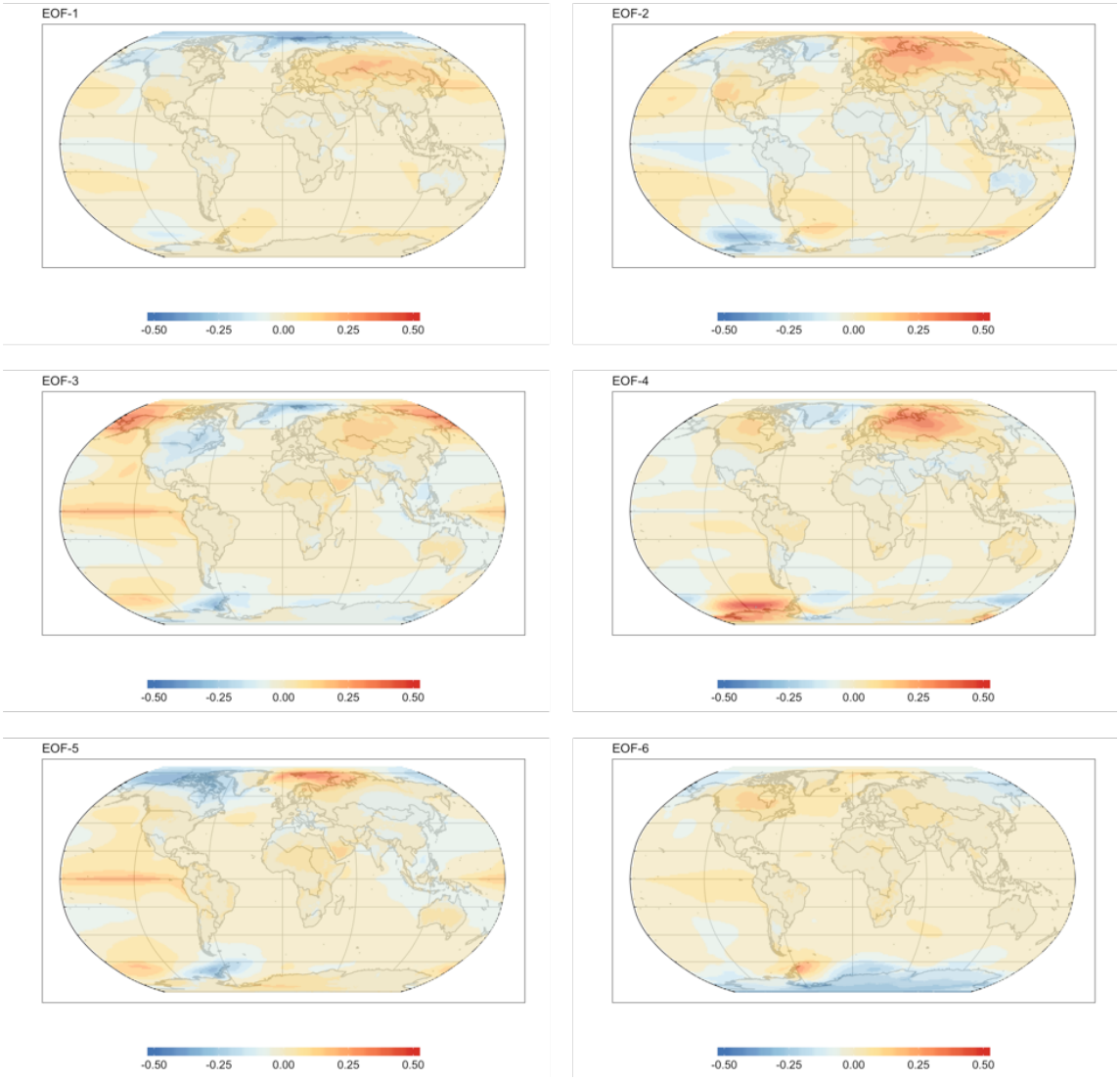

Figure 6. Spatial visualizations of the EOF1-6 basis functions. EOF grid cell values are scaled such that the magnitude of the largest value is 1. These components capture large-scale patterns of variability. EOFs 2, 3, and 5 all feature a temperature anomaly in the eastern Pacific. These same components 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.

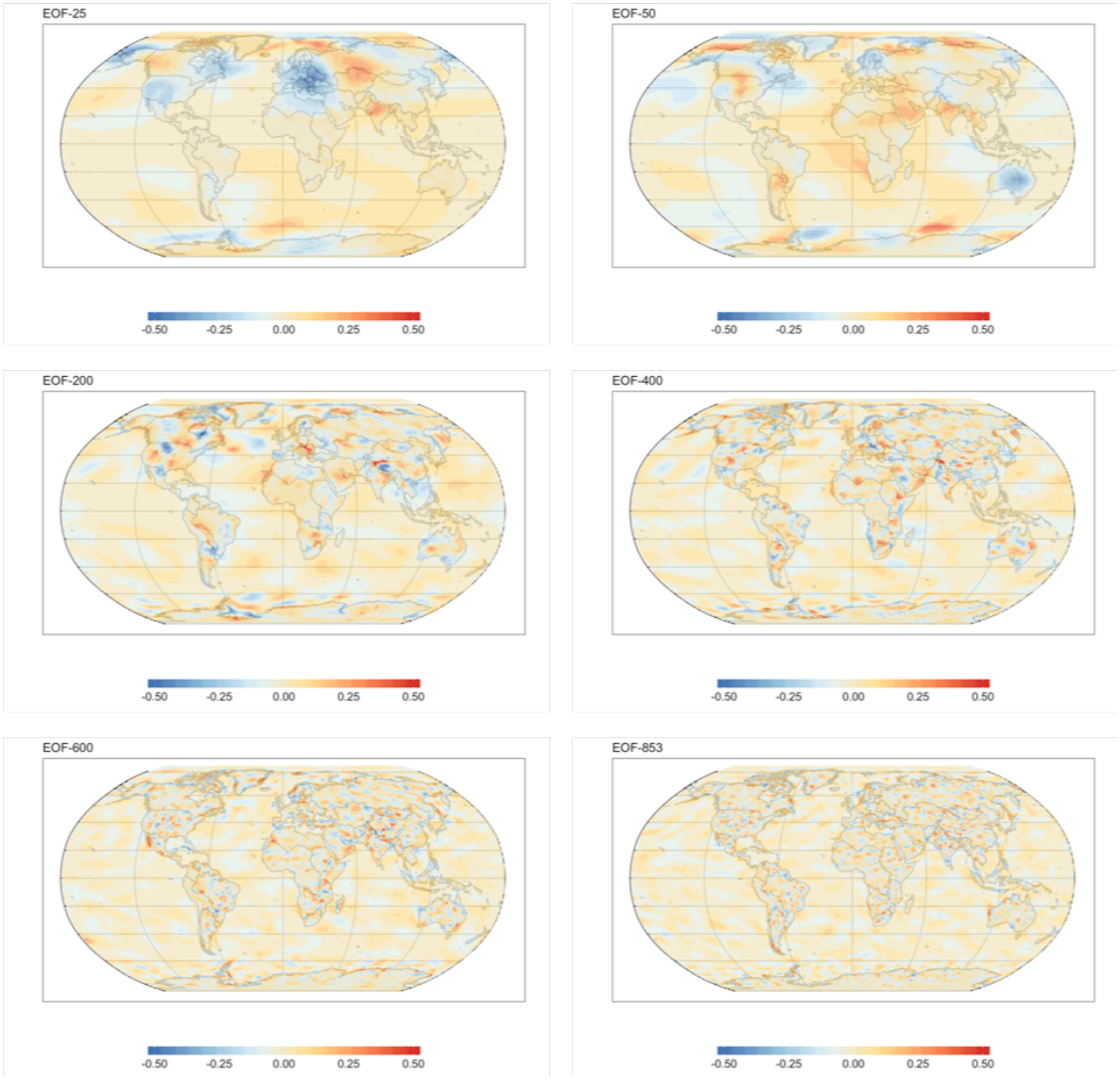

Figure 7. Spatial visualizations of higher EOF basis functions. ~~Note~~ EOF grid cell values are scaled such that the ~~decreasing~~ magnitude of the largest value is 1. The characteristic scale of temperature fluctuations decreases for ~~basis~~ functions later in the series. Thus, EOFs 25 and 50 show 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

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

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. 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~~ 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 ~~$\hat{x}_i$~~ $\hat{\boldsymbol{x}}_i$. 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 $\phi_i^\star$),

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

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 8, as our representative distribution for a *de minimis* effect size.

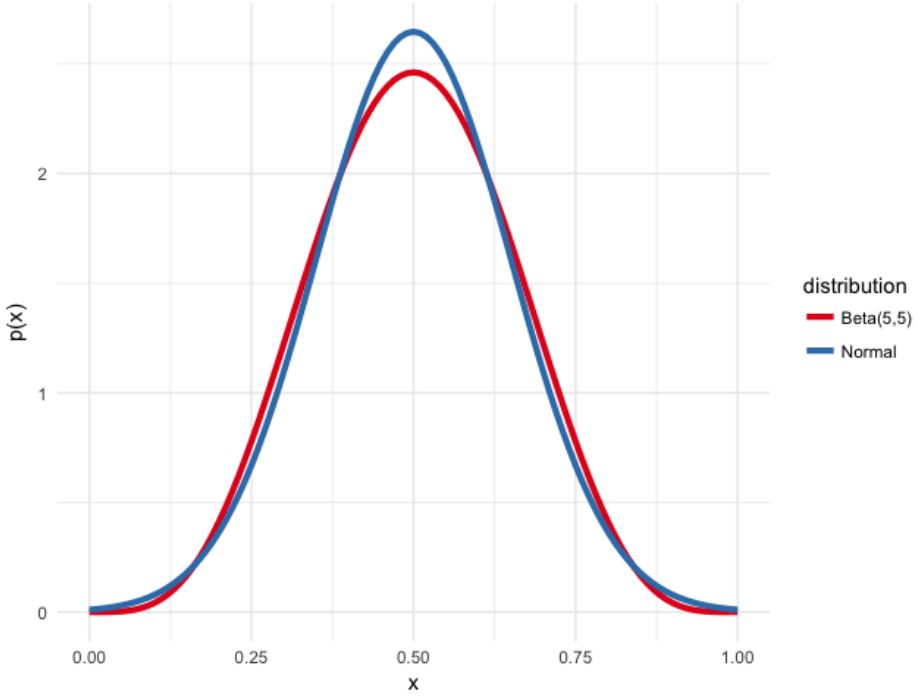

Figure 8. Comparison of the Beta(5,5) distribution and a Normal distribution with equal variance. The beta distribution is zero outside of the depicted range, while the normal distribution asymptotically approaches zero. Although the difference between these two distributions is small, the Shapiro-Wilk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

We used the Shapiro-Wilk ~~Test of Normality~~ 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 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.

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 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

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.