NIASRA
NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

University of Wollongong, Australia

Working Paper

3-16

Probabilistic Evaluation of Competing Climate Models

Amy Braverman, Snigdhansu Chatterjee, Megan Heyman, and Noel Cressie

National Institute for Applied Statistics Research Australia, University of Wollongong, Wollongong NSW 2522, Australia Phone +61 2 4221 5435, Fax +61 2 4221 4845.
Email: karink@uow.edu.au

Probabilistic Evaluation of Competing Climate Models

Amy Braverman*

Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA

Snigdhansu Chatterjee

University of Minnesota, Minneapolis, MN, USA

Megan Heyman

Rose-Hulman Institute of Technology, Terre Haute, IN, USA

Noel Cressie

University of Wollongong, Wollongong, Australia

**Corresponding author address:* Jet Propulsion Laboratory, Mail Stop 158-242, 4800 Oak Grove Drive, Pasadena, CA 91109-8099, USA.

E-mail: Amy.Braverman@jpl.nasa.gov

ABSTRACT

Climate models produce output over decades or longer at high spatial and temporal resolution. Starting values, boundary conditions, greenhouse gas emissions and so forth make the climate model an uncertain representation of the current climate system and, by implication, of the future climate system. Modern observational datasets offer opportunities for evaluation of competing climate models; in this article, we propose evaluation of competing climate models through probabilities. The probabilities are derived from summary statistics of climate model output and observational data, through a statistical resampling technique known as the Wild Scale-Enhanced Bootstrap. Here we compare monthly sequences of CMIP5 model output of average global near-surface temperature to similar sequences obtained from the well known HadCRUT4 data set. The summary statistics we choose come from working in the space of decorrelated and dimension-reduced wavelet space and regressing wavelet coefficients of model output on wavelet coefficients of observations. The dimension-reduced slope and intercept statistics are bootstrapped to allow a probability to be assigned to each model that reflects its output's compatibility with observations.

1. Introduction

Climate models are computational algorithms that model the climate system. They simulate many complex and inter-dependent processes, yielding global or regional fields that evolve from the past to the present and into the future. The models allow scientists to understand the consequences of different assumptions about both the physics of the climate system and forcings on it, including human influences. Climate models are also now viewed as decision-making tools because their projections of the future increasingly inform policy-making at the local, national, and international levels. The reliability of these future projections is central to both political and scientific debates about climate change.

Understanding climate and climate change is truly an international effort, with modeling centers from around the world contributing model runs for the most recent IPCC (Intergovernmental Panel on Climate Change) report. The diversity of scientific opinion reflected by these multiple runs, which use different initial conditions, parameterizations, and assumptions, is a key strength of this very democratic approach to science. However, it also leads to uncertainty because the results differ, and hence uncertainty quantification has become a critical issue in the interpretation of climate model output.

While the physical laws that underlie climate models are well understood, it is generally acknowledged that multiple sources of uncertainty continue to affect climate model projections. Broadly speaking, the sources of uncertainty that affect climate model simulations include natural climate variability at multiple scales, uncertainty in exogenous forcings such as anthropogenic greenhouse gas emissions, and uncertainty due to the models' abilities to represent the true physics of the climate system (Collins 2007).

Increasing computational power has made it possible to produce ensembles of runs under various controlled conditions, facilitating quantification of model uncertainty. Perturbed physics ensembles (PPEs) (Murphy et al. 2004; Deser et al. 2010) are created by running a single climate model multiple times with the model's parameters taking on different values for each trial. This allows quantification of the impact of uncertainty in these parameters on a model-by-model basis. Multi-model ensembles (MMEs; Tebaldi and Knutti (2007)) are constructed from single runs of each member of a collection of different climate models; they are aimed at quantifying so-called structural uncertainties, namely those due to "the numerical techniques used for solving the dynamical equations, the analytic form of parameterization schemes and the choices of inputs for fixed or varying boundary conditions" (Stocker et al. 2013).

There is by now a substantial literature on formal statistical modeling of climate model ensembles to produce probabilistic uncertainty estimates for future climate (Tebaldi et al. 2005; Rougier 2007; Smith et al. 2009; Stephenson et al. 2012; Rougier et al. 2013), and on the closely related topic of how to combine projections from its members (Min et al. 2007; Knutti et al. 2010). All these contributions rely on being able to specify a statistical model that describes the relationships among ensemble members' output and between those outputs and true climate. The latter is almost always achieved by comparing model output with observed data (Flato et al. 2013).

Typically, comparisons between climate model output and observed data are made on the basis of simple statistics, termed "metrics" in the literature (Gleckler et al. 2008). Observations are preprocessed by averaging across time and space to coincide with the resolution of climate model output (Teixeira et al. 2014), from which comparisons of means, medians, standard deviations, and correlations can be done straightforwardly. Results are often provided visually, using maps and other graphical devices, and they are not generally given probabilistic interpretations.

In this article, we propose a method for evaluating the fidelity of climate model runs to observed data that *does* produce a probabilistic measure of fidelity. For two time sequences, one produced by a climate model and one derived from observations, we test the null hypothesis that the "climate signals" (to be defined below) expressed by the two are the same. The probability under the null hypothesis that a given test statistic is equal to or more extreme than the observed value of the test statistic is called the p-value. A small p-value indicates incompatibility of the data with the null hypothesis (Wasserstein and Lazar 2016); in our case it indicates incompatibility of the climate model output with the observations.

Central to our approach is that climate signals are quantified in a spectral decomposition when a wavelet transform is applied to the time sequence. The level of agreement between the set of climate-signal wavelet coefficients derived from a climate model output and that of the corresponding observational sequence can be quantified by the intercept and slope obtained from a simple linear regression of the former on the latter. Our test statistic is constructed from these regression coefficients, and represents an important enhancement over current practice of using simple summary statistics that average over time to compare two series.

The null hypothesis we test is that the wavelet coefficients representing climate-scale behavior in the two series are the same. The null probability distribution that is required to perform this test is obtained using a new resampling technique that we call the Wild Scale-Enhanced (WiSE) Bootstrap. Thus, each model is assigned a p-value that can be used to weight the importance of the model in a multi-model ensemble. The reweighted p-values represent a probabilistic quantification of the uncertainty of the ensemble of models as judged by their compatibilities with the observations.

The remainder of this paper is organized as follows. In Section 2 we motivate our approach with a discussion of a probabilistic formalism for climate prediction, and we show the role of our contribution in facilitating it. Section 3 describes the WiSE Bootstrap and how it is used in this setting. In Section 4 we provide an end-to-end example of probabilistic climate model evaluation against observational data. We use monthly time sequences of global average near-surface temperature from a set of CMIP5 historical model runs for the period 1861–2005, which we compare to the HadCRUT4 monthly global average near-surface temperature data set. Conclusions follow in Section 5. There are two appendices: Appendix A gives a detailed, algorithmic description of our method, and Appendix B presents a simulation study that substantiates and quantifies the performance of our method on simulated data.

2. A probabilistic formalism for climate inference

This section explains how our methodology addresses the larger scientific objective of understanding and managing the uncertainties in climate model projections. We start from the probability model proposed in Rougier (2007) that relates model-generated and observed time sequences to that of true climate. Then, we identify the role of climate model output and how observational data can be used to to evaluate competing climate models and subsequently associate probabilities with them.

a. True climate and proxy time sequences

In what follows, we consider a single climate variable (e.g., global average near-surface temperature) whose true value is generically denoted as Y. Define $\boldsymbol{Y} = (Y_1,\ldots,Y_t,\ldots,Y_M)'$ to be a column vector of length M representing a sequence of values of Y through time. The vector $\boldsymbol{Y}$ can be partitioned as $\boldsymbol{Y} = \left(\boldsymbol{Y}_h,\boldsymbol{Y}_f\right)'$, where $\boldsymbol{Y}_h$ is the column vector of T components corresponding to the historical period, including the present, and $\boldsymbol{Y}_f$ is the column vector of $(M-T)$ components corresponding to the future.

Observations for the historical period are represented by the T-dimensional column vector $\boldsymbol{Z}_h$. In principle, statistical inference about climate (both historical and future) using observations, is based on the conditional distribution, $P(\boldsymbol{Y}|\boldsymbol{Z}_h)$, which is, via Bayes' Rule,

$$P(\boldsymbol{Y}|\boldsymbol{Z}_h) = \frac{P(\boldsymbol{Z}_h|\boldsymbol{Y})P(\boldsymbol{Y})}{P(\boldsymbol{Z}_h)}. \tag{1}$$

The right-hand side of Eq. (1) can be written as,

$$\begin{aligned}\frac{P(\boldsymbol{Z}_h|\boldsymbol{Y})P(\boldsymbol{Y})}{P(\boldsymbol{Z}_h)} &= \frac{P(\boldsymbol{Z}_h|\boldsymbol{Y}_h)}{P(\boldsymbol{Z}_h)}P(\boldsymbol{Y}_f|\boldsymbol{Y}_h)P(\boldsymbol{Y}_h),\\ &= \frac{P(\boldsymbol{Z}_h,\boldsymbol{Y}_h)}{P(\boldsymbol{Z}_h)}P(\boldsymbol{Y}_f|\boldsymbol{Y}_h),\\ &= P(\boldsymbol{Y}_h|\boldsymbol{Z}_h)P(\boldsymbol{Y}_f|\boldsymbol{Y}_h),\end{aligned} \tag{2}$$

where the first equality assumes, quite naturally, that historical data depend only on the historical climate, not the future climate.

The distribution $P(\boldsymbol{Y})$ is unknown, but the ensemble of climate model outputs provides us with a set of proxy sequences, $\{\boldsymbol{X}_l\}_{l=1}^{L}$, where L is the number of ensemble members. These are the result of L climate model runs; either runs of different models (a multi-model ensemble) or different runs of the same model with perturbed inputs (perturbed physics ensemble). A selection from the ensemble of climate model runs is represented by the vector $\boldsymbol{X}^{\dagger}$:

$$\boldsymbol{X}^{\dagger} = \sum_{l=1}^{L} 1_l \boldsymbol{X}_l, \tag{3}$$

where 1_l is an indicator taking value one if the l-th ensemble member is chosen, and zero otherwise.

We now break the problem into the two parts given by the right-hand side of Eq. (2). Write $\boldsymbol{X}_l = \left(\boldsymbol{X}_{lh}', \boldsymbol{X}_{lf}'\right)'$ and $\boldsymbol{X}^{\dagger} = \left(\boldsymbol{X}_h^{\dagger\prime}, \boldsymbol{X}_f^{\dagger\prime}\right)'$. We consider how well the probability distribution of $\boldsymbol{X}_f^{\dagger}|\boldsymbol{X}_h^{\dagger}$ represents the probability distribution of $\boldsymbol{Y}_f|\boldsymbol{Y}_h$, and how well the probability distribution of $\boldsymbol{X}_h^{\dagger}|\boldsymbol{Z}_h$ represents the probability distribution of $\boldsymbol{Y}_h|\boldsymbol{Z}_h$. Since our aim is to exploit the observations, and there are no observations of future climate, we focus on the second problem, which involves $\boldsymbol{X}_h^\dagger$ and $\boldsymbol{Z}_h$.

With respect to the historical period only, Eq. (3) becomes

$$\boldsymbol{X}_h^\dagger = \sum_{l=1}^{L} 1_l \boldsymbol{X}_{lh}. \tag{4}$$

Two sources of uncertainty contribute to uncertainty in $\boldsymbol{X}_h^\dagger$: randomness of the selection procedure represented by the random variables $\{1_l\}$, and the model uncertainty embodied by the random vectors $\{\boldsymbol{X}_{lh}\}$. We capture the model uncertainty by modeling each ensemble member $\boldsymbol{X}_l$ as a random vector, and hence $\boldsymbol{X}_{lh}$ is a time sequence covering the same historical period as $\boldsymbol{Z}_h$. We would like the distribution of the sequence $\boldsymbol{X}_h^\dagger|\boldsymbol{Z}_h$ to be a reasonable proxy for the distribution of the sequence $\boldsymbol{Y}_h|\boldsymbol{Z}_h$.

Our interest is in the evaluation of the members of the ensemble, and we shall reformulate this as specification of the marginal selection probabilities, $P(1_l = 1)$ for $l = 1, 2, \ldots, L$. This is the probabilistic uncertainty quantification referred to in Section 1. Assignment of the probabilities will be based on comparisons of $\boldsymbol{X}_{lh}$ to $\boldsymbol{Z}_h$, for $l = 1, \ldots, L$.

b. A statistical model for relating the proxy time sequence to true climate

Assume that the true historical sequence $\boldsymbol{Y}_h$, the l-th climate model's historical sequence $\boldsymbol{X}_{lh}$, and sequence of observations $\boldsymbol{Z}_h$, are related statistically as follows:

$$\boldsymbol{X}_{lh} = \boldsymbol{Y}_h + \boldsymbol{e}_{lh} \qquad \text{and} \qquad \boldsymbol{Z}_h = \boldsymbol{Y}_h + \boldsymbol{e}_{0h}, \tag{5}$$

where $\boldsymbol{e}_{lh}$ is the error of the l-th climate model sequence, and $\boldsymbol{e}_{0h}$ is an observational error term (Rougier 2007). Denote the joint distribution of $\boldsymbol{X}_{lh}$, $\boldsymbol{Y}_h$, and $\boldsymbol{Z}_h$ by $P(\boldsymbol{X}_{lh}, \boldsymbol{Y}_h, \boldsymbol{Z}_h)$; the conditional distribution, $P(\boldsymbol{X}_{lh}, \boldsymbol{Y}_h|\boldsymbol{Z}_h)$ quantifies the relationship between $\boldsymbol{X}_{lh}$, and $\boldsymbol{Y}_h$, conditional on the historical observations.

It remains to determine how the relationship between $\boldsymbol{X}_{lh}$ and $\boldsymbol{Y}_h$ can be quantified in order to model $P(1_l = 1)$. One obvious way would be through $\boldsymbol{D}_l = (\boldsymbol{X}_{lh} - \boldsymbol{Y}_h)$, and to assign $P(1_l = 1)$ proportional to the probability that $\boldsymbol{D}_l$ falls into some restricted region around the origin in high-dimensional space. Operationally, this would likely be difficult because of the high dimensionality and the ad hoc choice of a restricted region. The distance $D_l = ||\boldsymbol{D}_l||$ (or some weighted version) could be used instead, and we could assign $P(1_l = 1) \propto P(D_l \leq d)$, where d is a positive real number. However, taking the (possibly weighted) norm is a huge simplification that allows bad fidelity in one portion of the time sequence to offset good fidelity in another, which can lead to undesirable results. Moreover, these sequences exhibit temporal dependence, and so any methodology and its associated theory needs to incorporate this.

One way to account for temporal dependence is to transform the sequences so that the transformed values are decorrelated; in spectral analysis, this is sometimes called pre-whitening. In wavelet analysis, the Discrete Wavelet Transform (DWT) can be used:

$$\mathscr{C}_{\mathbf{X}} \equiv W\boldsymbol{X}, \tag{6}$$

where W is a square, orthonormal matrix (i.e., $W'W = \boldsymbol{I}$) that acts on a generic time sequence $\boldsymbol{X}$ resulting in the wavelet *coefficients* $\mathscr{C}_{\mathbf{X}}$ (Percival and Walden 2006). The choice of wavelet basis functions (father and mother wavelets) will determine the form of W.

In our analysis, we shall apply the same wavelet transform to detrended versions of $\{\boldsymbol{X}_{lh}\}$, $\boldsymbol{Y}_h$, and $\boldsymbol{Z}_h$; we work in the equivalent space of wavelet coefficients since those random quantities are decorrelated (Shen et al. 2002). Critically, our climate model evaluations are based on conditional distributions, $P(\mathscr{C}_{\mathbf{X}_{lh}}, \mathscr{C}_{\mathbf{Y}_h} | \mathscr{C}_{\mathbf{Z}_h})$, where $\mathscr{C}_{\mathbf{X}_{lh}}$, $\mathscr{C}_{\mathbf{Y}_h}$ and $\mathscr{C}_{\mathbf{Z}_h}$ denote coefficient vectors of $\boldsymbol{X}_{lh}$, $\boldsymbol{Y}_h$, and $\boldsymbol{Z}_h$, respectively.

We now establish some important notation for specifying the statistical models. Write $\boldsymbol{X}_{lh} = (X_{lh}(1),\ldots,X_{lh}(T))'$, $l = 1,\ldots,L$, and $\boldsymbol{Z}_h = (Z_h(1),\ldots,Z_h(T))'$. For the moment, assume that T is a power of two: $T = 2^J$. We model $X_{lh}(t)$ and $Z_h(t)$ as follows:

$$X_{lh}(t) = \gamma_{l0} + \gamma_{l1}t + \gamma_{l2}V_l(t/T) + \mu_l(t) + e_{lh}(t), \text{ for } t = 1,\ldots,T,\ l = 1,\ldots,L, \tag{7}$$

$$Z_h(t) = \gamma_{00} + \gamma_{01}t + \gamma_{02}V_0(t/T) + \mu_0(t) + e_{0h}(t), \text{ for } t = 1,\ldots,T, \tag{8}$$

where γ_{l0} and γ_{l1} are linear trend coefficients, and V_l are scaling coefficients, $l = 0,\ldots,L$. Note that the case $l = 0$ refers to quantities in the statistical model of the observations. In Eqs. (7) and (8),

$$\mu_l(t) = \sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1} \gamma_{ljk}W_{j,k}(t/T), \text{ for } l = 0,\ldots,L,\ t = 1,\ldots,T, \tag{9}$$

where $W_{j,k}(\cdot)$ is a fixed family of wavelet basis functions. The vectors of coefficients are

$$\mathscr{C}_{\boldsymbol{X}_{lh}} = \left(\gamma_{l0}, \gamma_{l1}, \gamma_{l2}, \gamma_{l00}, \ldots, \gamma_{l(J-1)(2^J-1)}\right)', \text{ for } l = 1,\ldots,L, \tag{10}$$

and

$$\mathscr{C}_{\boldsymbol{Z}_h} = \left(\gamma_{00}, \gamma_{01}, \gamma_{02}, \gamma_{000}, \ldots, \gamma_{0(J-1)(2^J-1)}\right)'. \tag{11}$$

Further, we assume that the noise terms, $e_{lh}(t)$ and $e_{0h}(t)$, are all mutually independent with means equal to zero but potentially unequal variances, $\mathrm{E}\left(e_{lh}^2(t)\right) = \sigma_l^2(t)$ and $\mathrm{E}\left(e_{0h}^2(t)\right) = \sigma_0^2(t)$.

The wavelet decomposition is a decorrelator, just like the usual Fourier spectral decomposition, but wavelets easily capture local behavior through functions that are of compact support, multi-resolutional, and translational within a resolution. The decorrelational aspect has proved particularly powerful for comparing two-dimensional spatial fields (Shen et al. 2002), and more recently Lin and Franzke (2015) showed that wavelets can capture multiresolution temporal structure in global average near-surface temperatures. Under the model (5), we would expect to see the wavelet coefficients associated with the l-th climate model, $\mathscr{C}_{\boldsymbol{X}_{lh}}$, track more or less closely those of the observations, $\mathscr{C}_{\boldsymbol{Z}_h}$.

c. Summary statistics that capture a relationship to the true climate

After applying W to the detrended versions of $\{\boldsymbol{X}_{lh}\}$ and $\boldsymbol{Z}_h$, we obtain the wavelet coefficients $\left\{\mathscr{C}_{\boldsymbol{X}_{lh}}\right\}$ and $\mathscr{C}_{\boldsymbol{Z}_h}$, respectively. The summary statistics that we shall use are based on a linear regression of $\mathscr{C}_{\boldsymbol{X}_{lh}}$ on $\mathscr{C}_{\boldsymbol{Z}_h}$. The wavelet coefficients are decorrelated and obtained from a linear transformation of the time sequence; hence, a plot of this regression line would all allow us to visualize the relationship between the output of a given climate model and the observations, without concern for misinterpretation due to temporal-dependence structures. Consider a generic climate model sequence, $\boldsymbol{X}_{lh}$; then the plot would ideally show that the coefficient pairs line up, with scatter, on a 45° line through the origin. When this does not happen, the obvious simple-linear-regression summary statistics, intercept $\hat{\alpha}_l$ and slope $\hat{\beta}_l$, express in wavelet space how "close" the climate model output comes to the noisy version of true climate provided by the observations. Thus, our evaluation of model l will be through comparison of $\left(\hat{\alpha}_l, \hat{\beta}_l\right)$ to the null value $(0,1)$ for each $l = 1, \ldots, L$.

Of course, we would prefer to compare $\{\boldsymbol{X}_{lh}\}$ directly to the true climate $\boldsymbol{Y}_h$, but a noisy version of it, $\boldsymbol{Z}_h$, is what we have. Hence, we denoise the observations to reveal the underlying climate signal. That is, we partition $\boldsymbol{Y}_h$ into a signal component, $\boldsymbol{Y}_h^s$, and a noise component, $\boldsymbol{Y}_h^n$, and we make a substitution of $\boldsymbol{X}_{lh}$ and $\boldsymbol{Z}_h$ in Eq. (5) with their wavelet coefficients, as follows.

$$\boldsymbol{Y}_h = \boldsymbol{Y}_h^s + \boldsymbol{Y}_h^n, \qquad \boldsymbol{X}_{lh} = \boldsymbol{Y}_h^s + \boldsymbol{Y}_h^n + \boldsymbol{e}_{lh}, \qquad \boldsymbol{Z}_h = \boldsymbol{Y}_h^s + \boldsymbol{Y}_h^n + \boldsymbol{e}_{0h}. \tag{12}$$

$$\mathscr{C}_{\boldsymbol{Y}_h} = \mathscr{C}_{\boldsymbol{Y}_h^s} + \mathscr{C}_{\boldsymbol{Y}_h^n}, \qquad \mathscr{C}_{\boldsymbol{X}_{lh}} = \mathscr{C}_{\boldsymbol{Y}_h^s} + \left(\mathscr{C}_{\boldsymbol{Y}_h^n} + \mathscr{C}_{\boldsymbol{e}_{lh}}\right), \qquad \mathscr{C}_{\boldsymbol{Z}_h} = \mathscr{C}_{\boldsymbol{Y}_h^s} + \left(\mathscr{C}_{\boldsymbol{Y}_h^n} + \mathscr{C}_{\boldsymbol{e}_{0h}}\right). \tag{13}$$

Here, $\mathscr{C}_{\boldsymbol{Y}_h^s}$, $\mathscr{C}_{\boldsymbol{Y}_h^n}$, $\mathscr{C}_{\boldsymbol{e}_{lh}}$, and $\mathscr{C}_{\boldsymbol{e}_{0h}}$ are the vectors of wavelet coefficients of $\boldsymbol{Y}_h^s$, $\boldsymbol{Y}_h^n$, $\boldsymbol{e}_{lh}$, and $\boldsymbol{e}_{0h}$, respectively. The terms in parentheses in Eq. (13) cannot be separately identified, so we consider them to be residual errors.

The key assumption that we shall make is that $\boldsymbol{Z}_h$ can be denoised to leave behind only the wavelet coefficients associated with climate signal, $\mathscr{C}_{\boldsymbol{Y}_h}$. Let $\check{J}$ be a constant, $\check{J} \leq J$, that specifies the number of coarse-scale wavelet-decomposition levels that define climate signal in the wavelet-level hierarchy. Let $\mathscr{S}(\mathscr{C}_{\boldsymbol{X}}, \check{J})$ be a smoothing function that operates on $\mathscr{C}_{\boldsymbol{X}}$ by setting elements corresponding to levels greater than $\check{J}$, to zero. So,

$$\begin{aligned}\mathscr{C}_{\boldsymbol{X}} &= \left(\gamma_{00}, \gamma_{01}, \ldots, \gamma_{(\check{J}-1)2^{(\check{J}-1)}}, \gamma_{\check{J}1}, \ldots, \gamma_{(J-1)(2^{J-1})}\right)', \\ \mathscr{S}(\mathscr{C}_{\boldsymbol{X}}, \check{J}) &= \left(\gamma_{00}, \gamma_{01}, \ldots, \gamma_{(\check{J}-1)2^{(\check{J}-1)}}, 0, \ldots\ldots\ldots, 0\right)',\end{aligned} \tag{14}$$

and the corresponding smoothed time sequence is $S(\boldsymbol{X}, \check{J}) = W' \mathscr{S}(\mathscr{C}_{\boldsymbol{X}}, \check{J})$. Our assumption is that after smoothing, $\mathscr{S}(\mathscr{C}_{\boldsymbol{Z}_h}, \check{J}) = \mathscr{C}_{\boldsymbol{Y}_h^s}$, the wavelet coefficients of the climate signal.

Climate model sequences $\{\boldsymbol{X}_{lh}\}$ can be evaluated according to how well their wavelet coefficients corresponding to levels $1, \ldots, \check{J}$, reproduce those of $\mathscr{C}_{\boldsymbol{Z}_h}$. Define $\mathscr{T}(\mathscr{C}_{\boldsymbol{X}}, \check{J})$ as a truncation operator that deletes all elements of $\mathscr{C}_{\boldsymbol{X}}$ that correspond to levels greater than $\check{J}$. Then,

$$\mathscr{T}\left(\mathscr{S}(\mathscr{C}_{\boldsymbol{X}}, \check{J}), \check{J}\right) = \left(\gamma_{00}, \gamma_{01}, \ldots, \gamma_{(\check{J}-1)2^{(\check{J}-1)}}\right)'. \tag{15}$$

Now the vectors $\{\boldsymbol{c}_l\}$ and $\boldsymbol{c}_0$ are defined as

$$\boldsymbol{c}_l = \mathscr{T}\left(\mathscr{S}(\mathscr{C}_{\boldsymbol{X}_{lh}}, \check{J}), \check{J}\right) \text{ for } l = 1, \ldots, L, \text{ and } \boldsymbol{c}_0 = \mathscr{T}\left(\mathscr{S}(\mathscr{C}_{\boldsymbol{Z}_h}, \check{J}), \check{J}\right). \tag{16}$$

For the l-th climate model, a low-dimensional summary of the agreement between $\boldsymbol{c}_l$ and $\boldsymbol{c}_0$ is motivated by simple linear regression. Define,

$$\bar{\gamma}_l = \left(\sum_{j=0}^{\check{J}-1}\sum_{k=0}^{2^j-1} 1\right)^{-1} \sum_{j=0}^{\check{J}-1}\sum_{k=0}^{2^j-1} \gamma_{ljk}, \quad l = 0, 1, \ldots, L,$$

$$\hat{\beta}_l = \left[\sum_{j=0}^{\check{J}-1}\sum_{k=0}^{2^j-1} \left(\gamma_{0jk} - \bar{\gamma}_0\right)^2\right]^{-1} \sum_{j=0}^{\check{J}-1}\sum_{k=0}^{2^j-1} \left(\gamma_{0jk} - \bar{\gamma}_0\right)\left(\gamma_{ljk} - \bar{\gamma}_l\right), \; l = 1, \ldots, L, \tag{17}$$

$$\hat{\alpha}_l = \bar{\gamma}_l - \hat{\beta}_l \bar{\gamma}_0, \quad l = 1, \ldots, L. \tag{18}$$

In what follows, we shall consider a test statistic based on $\hat{\alpha}_l$ and $\hat{\beta}_l$. It is crucial to obtain good estimates of the test statistic's variance under $H_0 : (\alpha_l, \beta_l) = (0, 1)$ against the alternative $H_A : (\alpha_l, \beta_l) \neq (0, 1)$; $l = 1, \ldots, L$. We shall obtain variance estimates using a technique we call the Wild Scaled-Enhanced Bootstrap (the WiSE bootstrap). Briefly, this method allows us to generate B "pseudo-realizations" of a time sequence from a single parent time sequence (under H_0) by perturbing the wavelet coefficients of the parent and inverting the wavelet transform. Then, for each pseudo-realization, indexed by b, we perform the wavelet decomposition and regression described above to obtain B resampled values, $\left\{\left(\hat{\alpha}^*_{lb}, \hat{\beta}^*_{lb}\right) : b = 1, \ldots, B\right\}$. The empirical variance of this bootstrap sample is an approximation to the sampling variance of $(\hat{\alpha}_l, \hat{\beta}_l)$; see Eqs. (35) and (36) below.

The quantile of $\left(\hat{\alpha}_l, \hat{\beta}_l\right)$ in the distribution of $\left\{\left(\hat{\alpha}^*_{lb}, \hat{\beta}^*_{lb}\right) : b = 1, \ldots, B\right\}$ is an empirical approximation to one minus the p-value of the test of the null hypothesis $H_0 : (\alpha_l, \beta_l) = (0, 1)$ under the conditions and assumptions described above. It is interpreted here as being proportional to a probability-scale measure of compatibility between the test statistic's value and how extreme it is under the null hypothesis. To emphasize this interpretation, we shall refer to these p-values as compatibility measures.

3. Statistical Methodology

In this section, we provide the details of our methodology for evaluating a set of climate models based on the statistical approach given in Section 2. There are four main steps: preprocessing, estimating the summary statistics, obtaining the null distribution of the summary statistics, and assignment of selection probabilities. From this point forward, all climate-variable sequences shall be understood to cover the historical period only, so for simplicity we drop the h subscript.

a. Preprocessing

Preprocessing is necessary for two reasons. First, it removes the effects of obvious, non-oscillatory components of the signals that are captured as trend in our models in Eqs. (7) and (8). Second, in order to apply the standard DWT software (e.g., R's wavethresh package due to Nason (2015)), the time sequences must have lengths that are powers of two.

Let N denote the original length of the time sequences, $\{\boldsymbol{X}_l\}$ and $\boldsymbol{Z}$, each indexed by $t = 1, \ldots, N$. To detrend, we fit simple linear regressions of $\boldsymbol{X}_l$ and $\boldsymbol{Z}$ on the vector $(1, \ldots, N)'$. This yields $\{(\hat{\gamma}_{l0}, \hat{\gamma}_{l1})\}$ and $(\hat{\gamma}_{00}, \hat{\gamma}_{01})$, respectively which are estimates of the trend intercepts and trend slopes for the climate model outputs ($l = 1, \ldots, L$) and the observations. Then the trend coefficients are obtained as follows: For $l = 1, \ldots, N$,

$$\bar{X}_l = N^{-1} \sum_{t=1}^{N} X_l(t), \tag{19}$$

$$\hat{\gamma}_{l1} = \left(\sum_{t=1}^{N} (t-(N+1)/2)^2 \right)^{-2} \sum_{t=1}^{N} (t-(N+1)/2) X_l(t), \tag{20}$$

$$\hat{\gamma}_{l0} = \bar{X} - \hat{\gamma}_{l1}(N+1)/2, \tag{21}$$

$$\bar{Z} = N^{-1} \sum_{t=1}^{N} Z(t), \tag{22}$$

$$\hat{\gamma}_{01} = \left(\sum_{t=1}^{N} (t-(N+1)/2)^2 \right)^{-2} \sum_{t=1}^{N} (t-(N+1)/2) Z(t), \tag{23}$$

$$\hat{\gamma}_{00} = \bar{Z} - \hat{\gamma}_{01}(N+1)/2. \tag{24}$$

Thus, the detrended series are:

$$\tilde{\tilde{X}}_l(t) = X_l - \hat{\gamma}_{l0} - \hat{\gamma}_{l1} t, \quad t = 1, \ldots, N,\ l = 1, \ldots, L, \tag{25}$$

$$\tilde{\tilde{Z}}(t) = Z(t) - \hat{\gamma}_{00} - \hat{\gamma}_{01} t, \quad t = 1, \ldots, N. \tag{26}$$

To prepare for the DWT, we pad $\tilde{\tilde{\boldsymbol{X}}}_l$ and $\tilde{\tilde{\boldsymbol{Z}}}$ so that they have lengths equal to $T = 2^{\lceil \log_2 N \rceil}$, where $\lceil \cdot \rceil$ is the ceiling function that returns the smallest integer greater than or equal to its argument. To do this, we reflect the appropriate subsequences of components at the beginning and end of each sequence. That is, let $m_1 = m_2 = \lceil (T-N)/2 \rceil$ if N is even, and if N is odd, let $m_1 = \lceil (T-N)/2 \rceil + 1$ and $m_2 = \lceil (T-N)/2 \rceil$. Then define the padded data as

$$\tilde{\boldsymbol{X}}_l = \left(\tilde{\tilde{X}}_{lm_1}, \ldots, \tilde{\tilde{X}}_{l2}, \tilde{\tilde{\boldsymbol{X}}}_l{}', \tilde{\tilde{X}}_{l(T-1)}, \ldots, \tilde{\tilde{X}}_{l(T-m_2)} \right)', \tag{27}$$

$$\tilde{\boldsymbol{Z}} = \left(\tilde{\tilde{Z}}_{m_1}, \ldots, \tilde{\tilde{Z}}_2, \tilde{\tilde{\boldsymbol{Z}}}', \tilde{\tilde{Z}}_{(T-1)}, \ldots, \tilde{\tilde{Z}}_{(T-m_2)} \right)'. \tag{28}$$

b. Estimating summary statistics

The second step is to obtain the simple-linear-regression summary statistics $(\hat{\alpha}_l, \hat{\beta}_l)$, $l = 1, \ldots, L$. We perform wavelet decompositions, with J levels, on $\tilde{\boldsymbol{X}}_l$ and $\tilde{\boldsymbol{Z}}_h$ using a common wavelet basis.

The model we use for the detrended, padded series with individual terms $\tilde{X}_l(t)$ and $\tilde{Z}(t)$ is:

$$\tilde{X}_l(t) = \gamma_{l2} V_l(t/T) + \mu_l(t) + e_l(t), \quad t = 1,\ldots,T, \quad l = 1,\ldots,L, \tag{29}$$

where

$$\mu_l(t) = \sum_{j=0}^{\check{J}} \sum_{k=0}^{2^j-1} \gamma_{ljk} W_{j,k}(t/T), \quad l = 1,\ldots,L; \tag{30}$$

and

$$\tilde{Z}(t) = \gamma_{02} V_0(t/T) + \mu_0(t) + e_0(t), \quad t = 1,\ldots,T, \tag{31}$$

where

$$\mu_0(t) = \sum_{j=0}^{J} \sum_{k=0}^{2^j-1} \gamma_{0jk} W_{j,k}(t/T). \tag{32}$$

Recall that $\check{J}$ is the wavelet decomposition level corresponding to the finest temporal scale deemed to represent climate signal. After performing the DWT on $\{\boldsymbol{X}_l : l = 1,\ldots,L\}$, and $\boldsymbol{Z}$, we obtain the wavelet coefficients

$$\begin{aligned}
\hat{\mathscr{C}}_{\tilde{\mathbf{X}}_l} &= \left(\hat{\gamma}_{l00}, \hat{\gamma}_{l01}, \ldots, \hat{\gamma}_{l(\check{J}-1)2^{(\check{J}-1)}}, \hat{\gamma}_{l\check{J}1}, \ldots, \hat{\gamma}_{l(J-1)(2^{J-1})}\right)', \\
\hat{\mathscr{C}}_{\tilde{\mathbf{Z}}} &= \left(\hat{\gamma}_{000}, \hat{\gamma}_{001}, \ldots, \hat{\gamma}_{0(\check{J}-1)2^{(\check{J}-1)}}, \hat{\gamma}_{0\check{J}1}, \ldots, \hat{\gamma}_{0(J-1)(2^{J-1})}\right)',
\end{aligned} \tag{33}$$

and we set

$$\begin{aligned}
\hat{\boldsymbol{c}}_l &= \mathscr{T}\left(\mathscr{S}(\hat{\mathscr{C}}_{\tilde{\mathbf{X}}_l}, \check{J}), \check{J}\right) = \left(\hat{\gamma}_{l00}, \hat{\gamma}_{l01}, \ldots, \hat{\gamma}_{l(\check{J}-1)2^{(\check{J}-1)}}\right), \\
\hat{\boldsymbol{c}}_0 &= \mathscr{T}\left(\mathscr{S}(\hat{\mathscr{C}}_{\tilde{\mathbf{Z}}}, \check{J}), \check{J}\right) = \left(\hat{\gamma}_{000}, \hat{\gamma}_{001}, \ldots, \hat{\gamma}_{0(\check{J}-1)2^{(\check{J}-1)}}\right).
\end{aligned} \tag{34}$$

Finally, summary statistics $\{(\hat{\alpha}_l, \hat{\beta}_l)\}$, $l = 1,\ldots,L$ in Eqs. (17) and (18) are computed from,

$$\begin{aligned}
\hat{\gamma}_l &= \left(\sum_{j=0}^{\check{J}-1} \sum_{k=0}^{2^j-1} 1\right)^{-1} \sum_{j=0}^{\check{J}-1} \sum_{k=0}^{2^j-1} \hat{\gamma}_{ljk}, \quad l = 0,1,\ldots,L, \\
\hat{\beta}_l &= \left[\sum_{j=0}^{\check{J}-1} \sum_{k=0}^{2^j-1} \left(\hat{\gamma}_{0jk} - \hat{\gamma}_0\right)^2\right]^{-1} \sum_{j=0}^{\check{J}-1} \sum_{k=0}^{2^j-1} \left(\hat{\gamma}_{0jk} - \hat{\gamma}_0\right)\left(\hat{\gamma}_{ljk} - \hat{\gamma}_l\right), \ l = 1,\ldots,L,
\end{aligned} \tag{35}$$

$$\hat{\alpha}_l = \hat{\gamma}_l - \hat{\beta}_l \hat{\gamma}_0, \quad l = 1,\ldots,L. \tag{36}$$

c. Obtaining the null distribution of the summary statistics

Under $H_{0l} : (\alpha_l, \beta_l) = (0, 1)$, the detrended series $\tilde{\boldsymbol{X}}_l$ and $\tilde{\boldsymbol{Z}}$ share the same climate signal. That is, $\boldsymbol{c}_l = \boldsymbol{c}_0$, or equivalently, $\{\mu_{lt}\} = \{\mu_{0t}\}$ in Eq. (9). To test H_{0l}, we will simulate the sampling distribution of $(\hat{\alpha}_l, \hat{\beta}_l)$ under this null hypothesis and assess the observed value, $(\hat{\alpha}_l, \hat{\beta}_l)$, against it. This results in a p-value, which we interpret as a measure of compatibility of the model output with the observations. Small values indicate incompatibility of the model output under consideration (Wasserstein and Lazar 2016). To do this, we create a collection of paired, resampled pseudo-series from the original, parent time sequences using a method based on the wild bootstrap (Wu 1986; Mammen 1993), under the assumption that the null hypothesis is true. For the lth model, denote the b-th pseudo-sequence pair by $\left\{\boldsymbol{X}^*_{lb}, \boldsymbol{Z}^*_b\right\}$ and the regression coefficients derived from it by $\left(\hat{\alpha}^*_{lb}, \hat{\beta}^*_{lb}\right)$. The empirical distribution of $\left\{\left(\hat{\alpha}^*_{lb}, \hat{\beta}^*_{lb}\right) : b = 1, 2, \ldots, B\right\}$ is then an estimate of the null distribution under H_{0l}. In Appendix A, we give the algorithmic details of this procedure, which we call the Wild Scale-Enhanced (WiSE) Bootstrap.

d. Computing compatibilities

We now compute compatibilities of the model outputs with the observations via tests of the null hypotheses, $H_{l0} : (\alpha_l, \beta_l) = (0, 1)$, for $l = 1, \ldots, L$. We use the test statistic,

$$Q_l = \begin{pmatrix} \hat{\alpha}_l & \hat{\beta}_l - 1 \end{pmatrix} \boldsymbol{K}^{-1} \begin{pmatrix} \hat{\alpha}_l \\ \hat{\beta}_l - 1 \end{pmatrix}, \tag{37}$$

where $\boldsymbol{K}$ is the bootstrap covariance matrix of $\left\{\hat{\alpha}^*_{bl}, \hat{\beta}^*_{bl} : b = 1, \ldots, B\right\}$, namely

$$\boldsymbol{K} = B^{-1} \begin{pmatrix} \sum_{b=1}^{B} (\hat{\alpha}^*_{bl} - \bar{\alpha}^*_l)^2 & \sum_{b=1}^{B} (\hat{\alpha}^*_{bl} - \bar{\alpha}^*_l)(\hat{\beta}^*_{bl} - \bar{\beta}^*) \\ \sum_{b=1}^{B} (\hat{\alpha}^*_{bl} - \bar{\alpha}^*_l)(\hat{\beta}^*_{bl} - \bar{\beta}^*) & \sum_{b=1}^{B} (\hat{\beta}^*_{bl} - \bar{\beta}^*_l)^2 \end{pmatrix}, \tag{38}$$

with $\bar{\alpha}_l^* = B^{-1}\sum_{b=1}^{B} \hat{\alpha}_{bl}^*$ and $\bar{\beta}_l^* = B^{-1}\sum_{b=1}^{B} \hat{\beta}_{bl}^*$. Finally, Q_l is evaluated relative to the bootstrap distribution based on

$$Q_{bl}^* = \begin{pmatrix} \hat{\alpha}_l^* & \hat{\beta}_l^* - 1 \end{pmatrix} \boldsymbol{K}^{-1} \begin{pmatrix} \hat{\alpha}_l^* \\ \hat{\beta}_l^* - 1 \end{pmatrix}, \text{ for } b = 1, \ldots, B. \tag{39}$$

Specifically, the p-value associated with our test is estimated by

$$P(Q_l^* > Q_l | H_{0l}) \equiv \frac{\#(Q_{bl}^* > Q_l)}{B}. \tag{40}$$

We call this the "compatibility" of model l's output time sequence with the observational sequence under the null hypothesis H_{0l} specified above (Wasserstein and Lazar 2016). In what follows, we assign a probability distribution to $\{1_l : l = 1, \ldots, L\}$ in Eq. (4) by making $P(1_l = 1)$ proportional to model l's p-value. Below we show that model averaging according to this probability assignment results in high compatibility with the observed time sequence.

4. Case study: Evaluating CMIP5 models using observations

In this section, we demonstrate the methodology described in Section 3 by applying it to the evaluation of monthly global average near-surface temperatures produced by 44 CMIP5 models. We evaluate these against a benchmark observational data set used in a similar comparison presented in the 2013 IPCC report specifically in Chapter 9, Evaluation of Climate Models, (Flato et al. 2013).

a. Data sources

In this subsection, we describe both the climate model outputs from CMIP5 and the global average near-surface temperature observations against which the CMIP5 climate models will be evaluated.

1) CLIMATE MODEL OUTPUT

The CMIP5 experiments are broadly divided into near-term and long-term, with the long-term experiments designed specifically for model evaluation (Taylor et al. 2012). One sub-category of long-term experiments are the so-called "historical" runs for which climate modeling centers have provided simulated time sequences from the mid-nineteenth though the early twenty-first centuries. These simulations start where pre-industrial control runs finish, and they are forced by both natural and anthropogenic conditions. Both simulated and observed time sequences exhibit variability due to these forcings and also due to internal variability, which is defined by Taylor et al. (2012) as "variations solely due to internal interactions within the complex nonlinear climate system." They go on to say, "A realistic climate model should exhibit internal variability with spatial and temporal structure like the observed" and caution that this does not mean there will be a one-to-one match between simulated and observed occurrences of specific events or patterns. In other words, statistical agreement is what matters in these comparisons, and this is precisely what our probability-based measure of compatibility focuses on.

We obtained time sequences of global monthly mean near-surface air temperature produced by 44 different CMIP5 models from the KNMI Data Explorer website (`https://climexp.knmi.nl/selectfield_cmip5.cgi?id=someone@somewhere`). Climate Data Explorer allows on-the-fly aggregation, averaging, and renormalization of data sets with a simple menu-driven interface. We selected all models for which the variable `tas` (near-surface air temperature) was available in the historical experiment, except for the GISS (Goddard Institute for Space Studies) models. For the GISS models, we limited our selection to those that were designated physics version 1 ("p1"), since they represent prescribed rather than calculated aerosol and ozone fields and thus more closely match what is done by the other centers for the historical experiment. The monthly global mean is expressed as an anomaly from the mean of the period 1960 – 1991, as in Flato et al. (2013). Where multiple runs (ensemble members) of the same model were available, we selected the ensemble mean. Most sequences cover the period 1850-2005, although some start as late as 1861 and some end as late as 2015. The common period that we shall use in our case study is January 1861 through November 2005. Table 1 lists the 44 models used in this study and the modeling centers that are responsible for them.

2) HadCRUT4 observations

Following Flato et al. (2013), we used the HadCRUT4 data set (Monice et al. 2012) as the observational time sequence. HadCRUT4 combines land, air, and sea-surface temperature data to produce a 100-member ensemble of monthly gridded surface temperature fields reaching back to 1850. Documentation for these data and an in-depth description of how they were produced can be found in Monice et al. (2012). As with the model simulations, we used the KNMI Climate Explorer to obtain the monthly global average near-surface temperature anomalies for the period 1850-2005, where the anomalies are computed relative to the average of the period 1960-1991. Our observational time sequence is computed from the median value of the 100 ensemble members' global average near-surface temperature value. Additional details can be found at `http://www.metoffice.gov.uk/ hadobs/hadcrut4/faq.html`.

b. Exploratory comparison

Figure 1 shows a sample of time sequence plots of the 44 CMIP5 model outputs, with the HadCRUT4 observations superimposed. All our sequences are truncated to the period January 1861 through November 2005, which is the period of intersection for all models and HadCRUT4. The figure is similar but not identical to Figure 9.8(a) in Flato et al. (2013) due to differences in normalization and masking. The HadCRUT4 values lie mostly inside the envelope defined by the 44 output sequences. Note that the spread among the model sequences appears to decrease over time, as does the variability of individual sequences including HadCRUT4. There are sharp increases in all the anomaly values starting in about 1961.

The cyclical nature of the these data is easier to see if their linear trends are removed. Figure 2 shows plots of $\tilde{\tilde{\boldsymbol{X}}}_l$ and $\tilde{\tilde{\boldsymbol{Z}}}_h$ computed in Eqs. (25) and (26), respectively. Both low-frequency and high-frequency components are evident.

c. Application of the WiSE bootstrap to comparison of CMIP5 model simulations and observed HadCRUT4

We shall now discuss how each of the four steps delineated in Sections 3a though 3d are applied in our analysis. We start from truncated sequences of length $N = 1739$ for the period January 1861 through November 2005, which is the longest period covered by all models' sequences simultaneously.

1) PREPROCESSING

As a first step, we removed the linear trend from each series by estimating the simple linear regression coefficients $(\hat{\gamma}_{l0}, \hat{\gamma}_{l1})$, $l = 1, \ldots, 44$ (Eqs. (19) through (21)) for the model sequences, and $(\hat{\gamma}_{00}, \hat{\gamma}_{01})$ (Eqs. (22) through (24)) for the observational sequence. The residuals from the regression lines defined by these estimated parameters are denoted by $\{\tilde{\tilde{\boldsymbol{X}}}_l,\ l = 1, \ldots, 44\}$ and $\tilde{\tilde{\boldsymbol{Z}}}$, respectively, as shown in Eqs. (25) and (26).

The second preprocessing step is to pad the sequences so that they have lengths equal to the next-largest power of two. In this case, we require sequences of length $T = 2048$, requiring that we pad the beginning of the series with 155 values and the end of the series with 154 values, as described in Eqs. (27) and (28). The padded values are the reflections of the first 155 and last 154 elements of the sequences, respectively. Denote the detrended, padded sequences by $\{\tilde{\boldsymbol{X}}_l : l = 1,\ldots,44\}$ and $\tilde{\boldsymbol{Z}}$, as in Eqs. (27) and (28).

2) Estimating summary statistics

Next, we obtain estimates of the slopes and intercepts of the regressions of the climate-scale wavelet coefficients of $\tilde{\boldsymbol{X}}_l$ on those of $\tilde{\boldsymbol{Z}}$. Formulas are given in Eqs. (35) and (36). We choose to set the threshold for distinguishing between climate-scale and noise at $\check{J} = 5$; see below for an explanation of this choice. That is, $\hat{\boldsymbol{c}}_l = \left(\hat{\gamma}_{l00}, \hat{\gamma}_{l01}, \ldots \hat{\gamma}_{l,5,32)}\right)'$, $l = 1,\ldots,44$, and $\hat{\boldsymbol{c}}_0 = \left(\hat{\gamma}_{000}, \hat{\gamma}_{001}, \ldots \hat{\gamma}_{0,5,32)}\right)'$; all these vectors are of length 64.

The choice of $\check{J}$ is important because it defines the set of temporal scales over which we shall evaluate agreement between models and observations. This may also be impacted by the choice of the wavelet basis; here we use the Daubechies Least Asymmetric wavelet family with eight vanishing moments (DB8). The choice of wavelet family was made after experimentation with this and other families, in the context of the simulation study reported in Appendix B. The choice of wavelet family did not affect our results significantly and so we used the DB8 family which was also used by Lin and Franzke (2015).

The threshold, $\check{J} = 5$ was chosen as follows. We examined the progressive reconstruction of the HadCRUT4 detrended and padded sequence as wavelet decomposition levels were added. The left panel of Figure 3 shows the original sequence in light gray, the reconstructed versions of the sequence using levels up to and including level 5 (thick black line) and up to and including level 6 (thin black line). The right panel of Figure 3 zooms in on the first 300 time points in order to highlight periodicity. The smoothed series using levels up to and including level 5 has a periodicity of roughly 180 months, while the smoothed series using levels up to and including level 6 has a periodicity of roughly 50 months. These correspond to cycles of about 15 and 4 years, respectively. While there is no hard-and-fast definition of climate time scale, we define it as corresponding to peroidicities of 15 years or more. That is sufficient to capture the Pacific Decadal Oscillation and the Atlantic Mulit-decadal Oscillation, though not the El Nino Southern Oscillation or the Madden-Julian Oscillation (Woods Hole Oceanographic Institution 2015).

We computed estimated regression coefficients using formulas given in in Section 3b. Results are shown in Table 2, and Figure 4 presents the same results in the form of a scatterplot of $\hat{\alpha}_l$ versus $\hat{\beta}_l$, with one symbol for each model. It is clear that that there is much less variability in the intercepts ($\hat{\alpha}_l$) than in the slopes ($\hat{\beta}_l$). Moreover, 35 of the 44 slope values are smaller than one, in some cases far below one. Slope coefficients less than one are characteristic of models for which climate-scale wavelet coefficients underestimate those of the observations.

3) Obtaining the null distribution of the summary statistics

To generate an approximation to the sampling distribution of $(\hat{\alpha}_l, \hat{\beta}_l)$ under $H_{0l} : (\alpha_l, \beta_l) = (0, 1)$, we follow the prescription of Section 3c. We fit a wavelet model using $J = 11$ to the detrended, padded, HadCRUT4 observational sequence, and we reconstruct the (detrended and padded) time sequence using levels $j = 0, 1, 2, 3, 4, 5$ (recall that $\check{J} = 5$). This smoothed sequence is $\hat{\boldsymbol{\mu}}_0 = (\hat{\mu}_0(1), \hat{\mu}_0(2), \ldots, \hat{\mu}_0(2048))'$, and it is the starting point for constructing a pair of bootstrap resamples; one for the lth climate model paired with one for the observations.

For the model's resample, we add both the model's trend and a pseudo-residual based on the model series, to $\hat{\boldsymbol{\mu}}_0$ (see Eq. (A1)). The model's pseudo-residual is the residual of the padded, raw model series relative to its level $\check{J} = 5$ smoothed version, multiplied by 1) independent standard normal random deviates, one for each time index, and 2) a scale-enhancement factor τ. For the observations' resample, we add both the observations' trend and a pseudo-residual based on the observational series, to $\hat{\mu}_0$ (see Eq. (A2)). The observations' pseudo-residual is the residual of the padded, raw observational series relative to it's level $\check{J} = 5$ smoothed version, multiplied by independent standard normal random deviates, one for each time index, and the same scale-enhancement factor τ. Finally, both sets of resamples are truncated at their beginning and end to remove the artificial values added by padding.

For $l = 1, \ldots, 44$, the result of the resampling procedure described above is a pair of bootstrap-resampled time sequences that share the same climate signal component and thus obey the conditions of the null hypothesis, $H_{0l} : (\alpha_l, \beta_l) = (0, 1)$. Figure 7 shows one of the resampled sequences plotted on the same graph. The resampled HadCRUT4 sequence is in green, with its smoothed version shown as the thick green line. The resampled model sequence (CCSM4 is used here as an example) is in blue, with its smoothed version shown as the thick blue line. This results in the bth instance, $\left(\hat{\alpha}^*_{bl}, \hat{\beta}^*_{bl}\right)$, obtained by regressing the wavelet coefficients corresponding to levels zero through five of the paired-resampled CCSM4 sequence on those of the corresponding pair-resampled HadCRUT4 sequence. We repeat the process to create a total of $B = 1000$ pairs, and perform regressions within each resampled pair as illustrated in Figure 6. This yields $\left\{(\hat{\alpha}^*_{bl}, \hat{\beta}^*_{bl}) : \; b = 1, \ldots, 1000\right\}$.

4) Computing compatibilities

The left panel of Figure 5 shows the scatterplot of the bootstrapped values $\left\{(\hat{\alpha}^*_{bl}, \hat{\beta}^*_{bl}) : \; b = 1, \ldots, 1000\right\}$ along with the actual value of $(\hat{\alpha}_l, \hat{\beta}_l)$ for the CCSM4-model-observation pair. Recall that the bootstrapped values were obtained under $H_{0l} : (\alpha_l, \beta_l) = (0, 1)$. To evaluate the l-th model, we require the proportion of resampled points that are further away, in terms of the scaled squared distance Q_l (given by Eq. (37)), from the point (0,1), than the red point at $(\hat{\alpha}_l, \hat{\beta}_l)$ is from $(0, 1)$. This is depicted in the right panel of Figure 5 by the proportion of the histogram that is to the right of the red vertical line, which is the bootstrapped p-value. Here, $Q_l = 3.155$, and 199 of the 1000 values of $\left\{Q^*_{bl}, b = 1, \ldots, 1000\right\}$ are greater than 3.155, leading to a WiSE bootstrap measure of compatibility for CCSM4 against HadCRUT4 of 0.199. Columns 3 and 8 of Table 3 show the compatibility values, p_l, for all 44 models, with non-zero values highlighted. Clearly, some models' compatibility values are quite high (e.g., CESM1-BCG), but many do not compare well to the HadCRUT4 observational sequence.

5) Results

In Table 3, ten of the 44 models have non-trivial compatibility values; $p_l \geq .001$. CESM1-BGC has the highest compatibility measure ($p_l = 0.241$), followed by GFDL-CM3 ($p_l = 0.200$) and CCSM4 ($p_l = 0.199$). CESM1 is a new version of CCSM4 and CESM1-BGC is a version that includes biogeochemistry. Next, HadGEM2-CC has compatibility measure $p_l = 0.137$, followed by two more CESM1 models: CESM1-FASTCHEM ($p_l = 0.089$) and CESM1-WACCM ($p_l = 0.078$). Rounding out the models with compatibilities greater than 0.001 are CanESM2 ($p_l = 0.015$), BNU-ESM ($p_l = 0.010$), and MPI-ESM-MR and NorESM1-ME, which both have $p_l = 0.001$. If we were to go beyond model evaluation and carry out significance testing of each of these hypothesis tests individually, we would reject the null hypotheses that CanESM2, BNU-ESM, MPI-ESM-MR, NorESM1-ME, and all models with $p_l < .001$ share the same HadCRUT4 climate-scale structure at the 0.05 significance level. We would not reject the null hypothesis for the models with $p_l > 0.05$. Testing a compound null hypothesis involving multiple models would require quantifying model dependence; in this article we consider model-ensemble members one-at-a-time.

Table 3 also compares our compatibility evaluation to two simple metrics sometimes used by the climate community: root mean squared error and correlation. The first statistic has been rescaled so that it is a number between zero and one, with higher values corresponding to better model agreement with observations. See Eq. (41) where we define *srmse*. Both simple metrics are computed using the original-length model simulations and the HadCRUT4 observational sequence. Define

$$srmse_l = 1 - \frac{rmse_l}{\max_k\{rmse_k\}}, \quad \text{where} \quad rmse_l = \sqrt{\sum_{t=1}^{N} [X_l(t) - Z(t)]^2}; \tag{41}$$

and define

$$corr_l = \frac{N^{-1}\sum_{t=1}^{N}[X_l(t) - \bar{X}_l)(Z(t) - \bar{Z})]}{\sqrt{N^{-1}\sum_{t=1}^{N}[X_l(t) - \bar{X}_l]^2}\sqrt{N^{-1}\sum_{t=1}^{N}[Z(t) - \bar{Z}]^2}}, \tag{42}$$

where $\bar{X}_l = N^{-1}\sum_{n=1}^{N} X_l(n)$ and $\bar{Z} = N^{-1}\sum_{n=1}^{N} Z(n)$.

Figures 8 and 9 show the same information as that contained in Table 3, but in the form of scatterplots of $srmse_l$ and $corr_l$ versus p_l, respectively. Both $srmse_l$ and $corr_l$ compute their measures of fitness on a time-point-by-time-point basis and then average over time. This could allow good performance in one part of sequence to offset poor performance in another part, making both metrics somewhat blunt instruments. Most of the values of $\{srmse_l : l = 1,\ldots,L\}$ lie between about 0.15 and 0.50, and only a few lie near the one-to-one line with $\{p_l : l = 1,\ldots,L\}$. Most of the values of $\{corr_l : l = 1,\ldots,L\}$ range between between 0.60 and 0.80, and there is no apparent relationship with $\{p_l : l = 1,\ldots,L\}$.

As usually implemented, the two traditional skill scores $srmse_l$ and $corr_l$ do not provide a probabilistic criterion against which to judge their magnitudes. This is an important shortcoming since science proceeds by evaluating the compatibility of theoretical predictions (e.g., climate model output) with observed evidence using discrepancies that are measured on a probability scale. WiSE bootstrap simulations using $srmse_l$ or $corr_l$ in place of our wavelet-based statistic could be performed. Nevertheless these two metrics require matching time points in the model and observational sequences, despite the assertion by Taylor et al. (2012) that no such one-to-one correspondence between model output and the observations should be expected. Our compatibility measure takes care of both problems, since it has a built-in probabilistic criterion, and owing to working in the wavelet domain it does not require a one-to-one correspondence of time points.

6) Multi-model averaging

Finally, it is sometimes observed that a time sequence of the multi-model means can outperform individual models. If the WiSE bootstrap compatibility values are accurate reflections of the fidelity of climate-model-generated time sequences to an observational benchmark, like HadCRUT4, then we might expect that modeling $P(1_l = 1)$ with $P_D(1_l = 1) \propto p_l$ and weight-averaging the models' output sequences according to these probabilities, would produce a multi-model time sequence that would perform well against HadCRUT4.

To explore this, we computed both a differentially weighted model mean, $\bar{\boldsymbol{X}}_D = (\bar{X}_D(1), \bar{X}_D(2), \ldots, \bar{X}_D(N))'$, with normalized weights,

$$P_D(1_l = 1) = \frac{p_l}{\sum_{k=1}^{44} p_k}, \tag{43}$$

and a uniformly weighted model mean, $\bar{\boldsymbol{X}}_U = (\bar{X}_U(1), \bar{X}_U(2), \ldots, \bar{X}_U(N))'$ with weights $P_U(1_l = 1) = 1/44$. That is,

$$\bar{X}_D(n) = \sum_{l=1}^{44} X_l(n) \left(\frac{p_l}{\sum_{k=1}^{44} p_k} \right) \quad \text{and} \quad \bar{X}_U(n) = \frac{1}{44} \sum_{l=1}^{44} X_l(n), \quad n = 1, 2, \ldots, 1739. \tag{44}$$

Figure 10 shows the climate-scale reconstructions of the HadCRUT4 observations and the two multi-model mean sequences. The WiSE bootstrap compatibilities for $\bar{\boldsymbol{X}}_D$ and $\bar{\boldsymbol{X}}_U$ are 0.519 and 0.000, respectively, demonstrating that using $P_D(1_l = 1) \propto p_l$ is vastly superior to using uniform weights in defining a multi-model ensemble. That is, the compatibilities $\{p_l : l = 1, \ldots, 44\}$ provided by the WiSE bootstrap imply a distribution on the ensemble of model sequences whose mean value is closer to the observational sequence.

5. Conclusions

In this final section, we draw some conclusions about probabilistic model evaluation, the assumptions and performance of the WiSE bootstrap method, and the performance of the CMIP5 climate models evaluated here. We close with a discussion of future work.

We have introduced a probabilistic method to determine the degree to which climate-scale temporal-dependence structures in an observational time sequence are reproduced by climate model-output time sequences. For a given climate model, the degree of agreement, or compatibility, is quantified by the p-value from a test of the null hypothesis that climate-scale temporal dependence is the same in both the observed and climate-model-generated time sequences. The p-value is the probability that a discrepancy as large or larger than that computed from the model-generated and observed sequences would be obtained, if the null hypothesis were true. In this context, a small p-value is indicative of a model-generated sequence that is incompatible with the climate signal embedded in the observed time sequence.

Of course, such conclusions are predicated on the assumptions of the hypothesis-testing framework. These include the underlying statistical models for the time sequences, how we define "climate scale" in the context of those models, the choice of test statistic, and how the sampling distribution of the test statistic is simulated under the null hypothesis. We have made certain choices in this work that we believe to be reasonable, but other choices are certainly possible. The choice of the wavelet-decomposition level that constitutes the boundary between climate signal and climate noise is particularly important, as experiments have shown that it can change the results substantially. Users of the WiSE bootstrap methodology are free to choose differently in accordance with their own scientific questions and opinions. In fact, one could test hypotheses about specific temporal scales based on wavelet coefficients corresponding to individual wavelet-decomposition levels. Other test statistics besides ours are also possible and likely useful, since slopes and intercepts from simple linear regression of wavelet coefficients provide only one of many possible test statistics.

Conclusions about the CMIP5 models themselves are as follows. Table 3 shows that, according to our analysis, CCSM-BGC, GFDL-CM3, and HadGEM2-CC are most compatible with the HadCRUT4 climate-scale temporal behavior, at least for global mean near-surface temperature on a monthly basis and "climate signal" defined as the five coarsest wavelet scales. Our numerical measure of how well these models do is given by the values of $\{p_l\}$. These values can be interpreted as measures of how compatible the actual time sequence from model l and the HadCRUT4 sequence are. For example, under the assumption that the NorESM-ME really does reproduce the climate signal in the HadCRUT4 observations, we would expect NorESM-ME to produce a time sequence as as unlike HadCRUT4 as the the one we obtained for this study, about one time in 1000. This implies low compatibility between the null hypothesis and the NorESM-ME output. Conversely, the CESM1-BGC model's time sequence has a compatibility of 241 times in 1000.

The WiSE bootstrap compatibility measures and the simple metrics based on root mean squared error and correlation tell different stories because they emphasize different things. Our method addresses whether the temporal dependence structure of the observations is reproduced by the climate model time sequences. Simple metrics based on averages over time do not address whether statistical structure in observations is preserved by climate model simulations, as called for in Taylor et al. (2012), nor do they provide probabilistic interpretation.

Finally, we are pursuing extensions in several areas. We believe that WiSE bootstrap could provide a basis for a weighting scheme for multi-model ensembles or perturbed physics ensembles. The probabilities $P_D(1_l = 1)$ can be used as marginal selection probabilities when drawing time sequences from an ensemble in order to form a mean sequence. However, joint probabilities would be required to define a probability structure that captures the dependence between models. A more complex multiple-testing framework will be required to assign joint probabilities rather than simple marginal probabilities.

There are natural extensions of the WiSE bootstrap to spatial and spatio-temporal contexts. Moving from one-dimensional to two-dimensional wavelets would allow us to use the same technology on maps as we have used here on time sequences. However, moving to three spatial dimensions, three spatial dimensions with time, and multivariate settings may not be straightforward since wavelets may not be suitable basis functions for these more complex problems. We are investigating the use of other kinds of basis functions in ongoing research.

Acknowledgments. This research was carried out partially at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. It was supported by NASA's Earth Science Data Records Uncertainty Analysis Program. In addition, Cressie's research was partially supported by a 2015–2017 Australian Research Council Grant, and Chatterjee's research was partially supported by the National Science Foundation (NSF) under grant # IIS-1029711. The authors would like to thank Huikyo Lee and Stephen Leroy for their thoughtful and thorough comments on this work.

APPENDIX A

The Wild Scale-Enhanced Bootstrap (WiSE bootstrap)

Starting with the original length-N sequences, $\boldsymbol{X}_l$ and $\boldsymbol{Z}$, we perform the following steps.

1. Set B (the number of trials), $J = \log_2 T$, $T = 2^{\lceil \log_2 N \rceil}$ (the length of the padded sequences), and $\check{J}$ (the number of levels in the wavelet decomposition that constitute climate signal).

2. Obtain $\tilde{\boldsymbol{X}}_l$ and $\tilde{\boldsymbol{Z}}$ by preprocessing on $\boldsymbol{X}_l$ and $\boldsymbol{Z}$ as specified in Section 3a. Retain the computed values of the trend coefficients, $(\hat{\gamma}_{l0}, \hat{\gamma}_{l1})$ and $(\hat{\gamma}_{00}, \hat{\gamma}_{01})$.

3. Perform the J-level wavelet decomposition on $\tilde{\boldsymbol{Z}}$ to obtain the set of wavelet coefficients $\hat{\boldsymbol{c}}_0 = \left(\hat{\gamma}_{000}, \hat{\gamma}_{001}, \ldots, \hat{\gamma}_{0(\check{J}-1)2^{(\check{J}-1)}}\right)$ as shown in Eq. (34).

4. Compute $\hat{\boldsymbol{\mu}}_0 = (\hat{\mu}_0(1), \hat{\mu}_0(2), \ldots, \hat{\mu}_0(T))'$ from $\tilde{\boldsymbol{Z}}$ as specified in Eq. (9):

$$\hat{\mu}_0(t) = \sum_{j=0}^{\check{J}-1} \sum_{k=0}^{2^j-1} \hat{\gamma}_{0jk} W_{j,k}(t/T), \quad t = 1, 2, \ldots, T.$$

5. Generate B pairs of pseudo-series, $\left\{\left(\boldsymbol{X}^*_{bl}, \boldsymbol{Z}^*_b\right) : b = 1, \ldots, B\right\}$. The bth pair contains a length-T pseudo-series derived from $\boldsymbol{X}_l$, denoted by $\boldsymbol{X}^*_{bl}$, and a length-T pseudo-series derived from $\boldsymbol{Z}$, denoted by $\boldsymbol{Z}^*_b$. To do this, create

$$\boldsymbol{X}^*_{bl} = (X^*_{bl}(1), \ldots, X^*_{bl}(T))', \text{ where } \quad X^*_{bl}(t) = \hat{\gamma}_{l0} + \hat{\gamma}_{l1} t + \hat{\mu}_0(t) + \tau U_b(t) R_l(t), \qquad \text{(A1)}$$

$$\boldsymbol{Z}^*_b = (Z^*_b(1), \ldots, Z^*_b(T))', \text{ where } \quad Z^*_b(t) = \hat{\gamma}_{00} + \hat{\gamma}_{01} t + \hat{\mu}_0(t) + \tau S_b(t) R_0(t), \qquad \text{(A2)}$$

for $U_b(t)$ and $S_b(t)$ mutually independent, standard normal random variables; $R_l(t) = (\tilde{X}_l(t) - \hat{\mu}_0(t))$, $R_0(t) = (\tilde{Z}(t) - \hat{\mu}_0(t))$; and τ is a constant that depends on T.

The scale-enhancement factor, τ, satisfies the conditions $\tau^2 \to \infty$, and $\tau^2/T \to 0$ as $T \to \infty$. This term is needed for asymptotic consistency of our results. Mathematical details are discussed in Chatterjee (2016). The factor τ has the same kind of effect that the choice of a smaller subsample or resample size has on the performance of subsampling for m-out-of-n bootstrap schemes (Politis and Romano 1994; Bickel et al. 1997; Shao 1996).

Here we use $\tau^2 = logT$, which satisfies the two conditions above. Note that the *same* values $\hat{\mu}_l(t) = \hat{\mu}_0(t)$ are used in Eqs. (A1) and (A2). Using the same values is required to enforce the null hypothesis.

6. For $b = 1, \ldots, B$, and a fixed l, obtain $\left(\hat{\alpha}^*_{lb}, \hat{\beta}^*_{lb}\right)$ from $\boldsymbol{X}^*_{bl}$ and $\boldsymbol{Z}^*_b$ as follows.

    (a) Obtain $\tilde{\boldsymbol{X}}^*_{bl}$ and $\tilde{\boldsymbol{Z}}^*_b$ by preprocessing $\boldsymbol{X}^*_{bl}$ and $\boldsymbol{Z}^*_b$ as specified in Section 3a.

    (b) Perform wavelet decompositions on $\tilde{\boldsymbol{X}}^*_{bl}$ and $\tilde{\boldsymbol{Z}}^*_b$ to obtain wavelet coefficients $\hat{\boldsymbol{c}}^*_{bl} = \left(\hat{\gamma}^*_{bl00}, \hat{\gamma}^*_{bl01}, \ldots, \hat{\gamma}^*_{bl(\check{J}-1)2^{(\check{J}-1)}}\right)$ and $\hat{\boldsymbol{c}}^*_{b0} = \left(\hat{\gamma}^*_{b000}, \hat{\gamma}^*_{b001}, \ldots, \hat{\gamma}^*_{b0(\check{J}-1)2^{(\check{J}-1)}}\right)$, as shown in Eq. (34). Recall that $\check{J} \le J$ is the number of wavelet decomposition levels that define the climate signal in the time sequences.

    (c) Regress the elements of $\hat{\boldsymbol{c}}^*_{bl}$ on the corresponding elements of $\hat{\boldsymbol{c}}^*_{b0}$ using simple linear regression. Define

$$
\hat{\bar{\gamma}}^*_{bl} = \left(\sum_{j=0}^{\check{J}-1} \sum_{k=0}^{2^j-1} 1 \right)^{-1} \sum_{j=0}^{\check{J}-1} \sum_{k=0}^{2^j-1} \hat{\gamma}^*_{bljk},
$$

$$
\hat{\beta}^*_{bl} = \left[\sum_{j=0}^{\check{J}-1} \sum_{k=0}^{2^j-1} \left(\hat{\gamma}^*_{b0jk} - \hat{\bar{\gamma}}^*_{b0} \right)^2 \right]^{-1} \sum_{j=0}^{\check{J}-1} \sum_{k=0}^{2^j-1} \left(\hat{\gamma}^*_{b0jk} - \hat{\bar{\gamma}}^*_{b0} \right) \left(\hat{\gamma}^*_{bljk} - \hat{\bar{\gamma}}^*_{bl} \right), \tag{A3}
$$

$$
\hat{\alpha}^*_{bl} = \hat{\bar{\gamma}}^*_{bl} - \hat{\beta}^*_{bl} \hat{\bar{\gamma}}^*_{b0}, \tag{A4}
$$

    (d) The set $\left\{ \left(\hat{\alpha}^*_{bl}, \hat{\beta}^*_{bl}\right) : b = 1, 2, \ldots, B \right\}$, gives an approximation to the null distribution of $\left(\hat{\alpha}_l, \hat{\beta}_l\right)$ under $H_0 : (\alpha_l, \beta_l) = (0, 1)$.

APPENDIX B

Simulation Study

We conducted a simulation study to understand the performance of our proposed hypothesis testing method. We generated data from the two processes, $Y_{1t} = S_{1t} + \varepsilon_{1t}$ and $Y_{2t} = S_{2t} + \varepsilon_{2t}$, for $t =$

$1,\ldots,2^J$. Here, the first series Y_{1t} acts as the "observations" and the second series, Y_{2t}, acts as the "model output". In all the cases we consider below, $\{\varepsilon_{1t}\}$ and $\{\varepsilon_{2t}\}$ are mutually independent and identically distributed as $N(0,V)$. However, we have conducted studies with heteroskedastic noise, and the results do not change in substance from those reported below.

The signal components of both the processes Y_{1t} and Y_{2t} satisfy the framework we adopt in this paper, namely

$$S_{1t} = \sum_{j=0}^{J_0-1}\sum_{k=0}^{2^j-1}\gamma_{1jk}A_{j,k}(t), \tag{B1}$$

$$S_{2t} = \sum_{j=0}^{J_0-1}\sum_{k=0}^{2^j-1}\gamma_{2jk}A_{j,k}(t), \tag{B2}$$

where $\{A_{j,k}\}$ are a fixed set of wavelet basis functions, and $\gamma_{2jk} = \alpha+\beta\gamma_{1jk}$, which directly models the sort of relation we have in mind between model output and observations.

This simulation is achieved by starting with a series $\{X_t, t=1,\ldots,N\}$, and obtaining a wavelet decomposition of it; then consider the first $\check{J}$ coarse levels from it as defining $\{S_{1t}\}$ in the temporal domain. We constructed $\{X_t\}$ to follow an $AR(1)$ process that imitated the observed HadCRUT4 temperature data series. Note that the actual structure of the series $\{X_t\}$ is not relevant, since it is merely used to elicit a few coarse wavelet coefficients for the series $\{S_{1t}\}$. For the series $\{S_{2t}\}$, we used the relation given above for the wavelet basis functions $\{A_{j,k}(\cdot)\}$ and reconstructed $\{S_{2t}\}$ from them.

In all the scenarios described above, we obtained the p-value of the test $H_0 : (\alpha,\beta) = (0,1)$ against the alternative hypothesis $H_1 : (\alpha,\beta) \neq (0,1)$. Different values of β were used, and to simplify the study we chose $\alpha = 0$. The constants used for the simulations are given as follows.

1. We consider sample sizes, $N \in \{100,300,600,1000\}$.

2. We consider noise variances, $V \in \{0.01,0.2,0.5\}$.

3. We consider true scales for the coarse wavelet signal, $\check{J} \in \{1,3,5\}$.

4. We consider the resample size (bootstrap sample size), $B = 500$.

5. Each of the above scenarios is independently replicated $R = 200$ times.

We have tried other variations of 1.–5. that are not reported below. They include cases where the noise has unequal variances (depending on t); unequal variances for the observations, $\{Y_{1t}\}$ and the model output, $\{Y_{2t}\}$; other values of N in order to evaluate the effect of the padding; other values of $\check{J}$; other values of (α, β), and both larger and smaller values of B. Also, we used multiple ways of generating the signal component $\{S_{1t}\}$, that is, multiple ways of obtaining the initial time sequence $\{X_t\}$. We included trends, both in the $\{Y_{1t}\}$ and the $\{Y_{2t}\}$ series. All of these led to results that mimic the results below.

To illustrate the power-function properties of the proposed hypothesis tests, we fixed the size (maximum allowable probability of type 1 error, which is the probability of rejecting the null hypothesis when it is true) at 0.05, and we studied the power as β varied from 0.5 to 1.5. The power of a hypothesis test is defined as the probability of rejecting the null hypothesis when it is false; thus, a higher power is desirable.

In Figure B1, we present a selection of the power curves that we obtained from our simulations. Here, the figures in the left panels correspond to the sample size $N = 600$, and those in the right panels are for $N = 1000$. The top two panels, (a) and (b), use noise variance $V = 0.01$ and number of coarse wavelet levels $\check{J} = 3$; the middle two panels, (c) and (d), retain the same noise variance but use $\check{J} = 5$ wavelet levels; and the bottom two panels, (e) and (f), keep $\check{J} = 5$ but increase the noise variance to $V = 0.2$. In all the figures, the red horizontal line is drawn at the probabilities 0.05.

The power curves illustrate that our proposed method performs as expected in all simulation scenarios. First, corresponding to the null hypothesis regime of $\beta = 1$ in the center of each of the figures, the probability of rejection is lower than 0.05; thus we maintain the specified size properties. In all situations, it seems that our test is slightly conservative in the sense that the actual probability of rejecting a true null hypothesis is lower than 0.05. As $|\beta - 1|$ increases and the signal for the alternative hypothesis becomes stronger, the power curves increase (sometimes quite steeply) to 1.

The figures show that (*i*) power increases with sample size, when we compare the right panels with $N = 1000$ with the corresponding left panels for $N = 600$; (*ii*) the power increases with the signal, quantified by the increase in $\check{J}$, when we compare panel (a) with panel (c) or panel (b) with panel (d); and (*iii*) the power decreases with increased noise variance V, when we compare panel (c) with panel (e) or panel (d) with panel (f).

References

Bickel, P., F. Gotze, and W. van Zwet, 1997: Resampling fewer than n observations: gains, losses and remedies for losses. *Statistica Sinica*, **7**, 1–31.

Chatterjee, S., 2016: The WiSE bootstrap method for hypothesis testing on related slope parameters. Tech. rep., School of Statistics, University of Minnesota.

Collins, M., 2007: Ensembles and probabilities: a new era in the prediction of climate change. *Philosophical Transactions of the Royal Society, Series A*, **365**, 1957–1970, doi:10.1098/rsta.2007.2068.

Deser, C., A. Phillips, V. Bourdette, and H. Teng, 2010: Uncertainty in climate change projections: the role of internal variability. *Climate Dynamics*, **38**, 527–546, doi:10.1007/s10584-006-9156-9.

Flato, G., and Coauthors, 2013: Evaluation of climate models. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. Midgley, Eds., Cambridge University Press.

Gleckler, P., K. Taylor, and C. Doutriaux, 2008: Performance metrics for climate models. *Geophysical Research Letters*, **113**, doi:10.1029/2007JD008972.

Knutti, R., R. Furrer, C. Tebaldi, J. Cermak, and G. Meehl, 2010: Challenges in combining projections from multiple climate models. *Journal of Climate*, **23**, 2739–2758, doi:10.1175/2009JCLI3361.1.

Lin, Y., and C. Franzke, 2015: Scale-dependency of the global mean surface temperature trend and its implication for the recent hiatus of global warming. *Scientific Reports*, **5**, doi:10.1038/srep12971.

Mammen, E., 1993: Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics*, **21**, 255–285.

Min, S.-K., D. Simonis, and A. Hense, 2007: Probabilistic climate change predictions applying Bayesian model averaging. *Philosophical Transactions of the Royal Society, Series A*, **365**, 2103–2116, doi:10.1098/rsta.2007.2070.

Monice, C., J. Kennedy, N. N.A. Rayner, and P. Jones, 2012: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 dataset. *Journal of Geophysical Research*, **117**, doi:10.1029/2011JD017187.

Murphy, J., D. Sexton, D. Barnett, G. Jones, M. Webb, M. Collins, and D. Stainforth, 2004: Quanfification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, **430**, 768–772, doi:10.1038/nature02771.

Nason, G., 2015: *Package "wavethresh"*. Https://cran.r-project.org/web/packages/wavethresh/wavethresh.pdf.

Percival, D., and A. Walden, 2006: *Wavelet Methods for Time Series Analysis*. Cambridge University Press.

Politis, D., and J. Romano, 1994: Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, **22 (4)**, 2031–2050.

Rougier, J., 2007: Probabilistic inference for future climate using an ensemble of climate model evaluations. *Climatic Change*, **81**, 247–264, doi:10.1007/s10584-006-9156-9.

Rougier, J., M. Goldstein, and L. House, 2013: Second-order exchangeability analysis for multimodel ensembles of climate models. *Journal of the American Statistical Association*, **104**, 97–116.

Shao, J., 1996: Bootstrap model selection. *Journal of the American Statistical Association*, **91**, 655–665.

Shen, X., H. Huang, and N. Cressie, 2002: Nonparametric hypothesis testing for a spatial signal. *Journal of the American Statistical Association*, **97 (460)**, 1122–1140.

Smith, R., C. Tebaldi, D. Nychka, and L. Mearns, 2009: Bayesian modeling of uncertainty in ensembles of climate models. *Journal of the American Statistical Association*, **104 (485)**, 97–116, doi:10.1198/jasa.2009.0007.

Stephenson, D., M. Collins, J. Rougier, and R. Chandler, 2012: Statistical problems in the probabilistic prediction of climate change. *Environmetrics*, doi:10.1002/env.2153.

Stocker, T. F., and Coauthors, Eds., 2013: *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press.

Taylor, K., R. Stouffer, and G. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, 485–498, doi:10.1175/BAMS-D-11-00094.1.

Tebaldi, C., and R. Knutti, 2007: The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society, Series A*, **365**, 2053–2075, doi: 0.1098/rsta.2007.2076.

Tebaldi, C., R. Smith, D. Nychka, and L. Mearns, 2005: Quantifying uncertainty in projections of regional climate change: a Bayesian approach to the analysis of multi-model ensembles. *Journal of Climate*, **18**, 1524–1540, doi:10.1175/JCLI3363.1.

Teixeira, J., D. Waliser, R. Ferraro, P. Gleckler, T. Lee, and G. Potter, 2014: Satellite observations for CMIP5, the genesis of obs4mips. *Bulletin of the American Meteorological Society*, doi: 10.1175/BAMS-D-12-00204.1.

Wasserstein, R., and N. Lazar, 2016: The ASA's statement on *p*-values: context, process, and purpose. *The American Statistician*, doi:10.1080/00031305.2016.1154108.

Woods Hole Oceanographic Institution, 2015: `https://www.whoi.edu/main/topic/el-nino-other-oscillations`.

Wu, C., 1986: Jacknife, bootstrap, and other resampling methods in regression analysis. *The Annals of Statistics*, **14**, 1261–1295.

LIST OF TABLES

TABLE 1. 44 CMIP5 models used in this study.

l	*Model*	*Center*	*l*	*Model*	*Center*
1	ACCESS1-0	CSIRO-BOM (Australia)	23	GFDL-ESM2M	GFDL (USA)
2	ACCESS1-3	CSIRO-BOM (Australia)	24	GISS-E2-H p1	NASA GISS (USA)
3	BCC-CSM-1	Beijing Climate Center (PRC)	25	GISS-E2-H-CC p1	NASA GISS (USA)
4	BCC-CSM-1-M	Beijing Climate Center (PRC)	26	GISS-E2-R p1	NASA GISS (USA)
5	BNU-ESM	Beijing Normal Univ. (PRC)	27	GISS-E2-R-CC p1	NASA GISS (USA)
6	CanSM2	CCCMA (Canada)	28	HadGEM2-AO	NIMR/KMA (UK/Korea)
7	CCSM4	NCAR (USA)	29	HadGEM2-CC	MOHC/INPE (UK/Brazil)
8	CESM1-BGC	NCAR/DOE/NSF (USA)	30	HadGEM2-ES	MOHC/INPE (UK/Brazil)
9	CESM1-CAM5	NCAR/DOE/NSF (USA)	31	INMCM4	INM (Russia)
10	CESM1-CAM5-1-FV2	NCAR/DOE/NSF (USA)	32	IPSL-CM5A-LR	IPSL (France)
11	CESM1-FASTCHEM	NCAR/DOE/NSF (USA)	33	IPSL-CM5A-MR	IPSL (France)
12	CESM1-WACCM	NCAR/DOE/NSF (USA)	34	IPSL-CM5B-LR	IPSL (France)
13	CMCC-CESM	CMCC (Italy)	35	MIROC-ESM	MIROC (Japan)
14	CMCC-CM	CMCC (Italy)	36	MIROC-ESM-CHEM	MIROC (Japan)
15	CMCC-CMS	CMCC (Italy)	37	MIROC5	MIROC (Japan)
16	CNRM-CM5	CNRM (France)	38	MPI-ESM-LR	MPI (Germany)
17	CSIRO-Mk3-6-0	CSIRO (Australia)	39	MPI-ESM-MR	MPI (Germany)
18	EC-EARTH	EC-EARTH Consortium (Europe)	40	MPI-ESM-P	MPI (Germany)
19	FGOALS-g2	LASG (PRC)	41	MRI-CGM3	MRI (Japan)
20	FIO-ESM	FIO (PRC)	42	MRI-ESM1	MRI (Japan)
21	GFDL-CM3	GFDL (USA)	43	NorESM1-M	NCC (Norway)
22	GFDL-ESM2G	GFDL (USA)	44	NorESM1-ME	NCC (Norway)

TABLE 2. Intercept and slope estimates obtained from regressions of the elements of $\hat{\boldsymbol{c}}_l$ on the corresponding elements of $\hat{\boldsymbol{c}}_0$, for $l = 1, \ldots, 44$.

l	*Model*	$\hat{\beta}_l$	$\hat{\alpha}_l$	l	*Model*	$\hat{\beta}_l$	$\hat{\alpha}_l$
1	ACCESS1-0	0.764	-0.083	23	GFDL-ESM2M	0.691	0.005
2	ACCESS1-3	0.607	-0.064	24	GISS-E2-H p1	0.647	-0.018
3	BCC-CSM-1	1.161	-0.017	25	GISS-E2-H-CC p1	0.795	0.013
4	BCC-CSM-1-M	0.747	0.002	26	GISS-E2-R p1	0.697	-0.013
5	BNU-ESM	1.184	-0.024	27	GISS-E2-R-CC p1	0.647	-0.031
6	CanESM2	1.067	0.026	28	HadGEM2-AO	1.129	-0.103
7	CCSM4	1.044	0.018	29	HadGEM2-CC	0.915	0.006
8	CESM1-BGC	1.074	0.033	30	HadGEM2-ES	0.713	-0.036
9	CESM1-CAM5	0.898	0.068	31	INMCM4	0.485	-0.004
10	CESM1-CAM5-1-FV2	0.777	0.009	32	IPSL-CM5A-LR	1.093	-0.04
11	CESM1-FASTCHEM	1.069	-0.03	33	IPSL-CM5A-MR	0.86	-0.024
12	CESM1-WACCM	1.06	0.094	34	IPSL-CM5B-LR	0.519	0.019
13	CMCC-CESM	0.59	0.086	35	MIROC-ESM	0.689	0.004
14	CMCC-CM	0.658	0.034	36	MIROC-ESM-CHEM	0.608	0.003
15	CMCC-CMS	0.765	0.036	37	MIROC5	0.669	0.005
16	CNRM-CM5	0.855	-0.03	38	MPI-ESM-LR	0.84	-0.038
17	CSIRO-Mk3-6-0	0.592	-0.056	39	MPI-ESM-MR	0.92	-0.031
18	EC-EARTH	0.764	-0.018	40	MPI-ESM-P	0.806	0.02
19	FGOALS-g2	0.707	0.032	41	MRI-CGM3	0.445	0.024
20	FIO-ESM	0.672	0.042	42	MRI-ESM1	0.575	0.024
21	GFDL-CM3	0.961	-0.03	43	NorESM1-M	0.705	0.02
22	GFDL-ESM2G	0.682	-0.043	44	NorESM1-ME	0.844	-0.099

TABLE 3. WiSE bootstrap compatibilities, p_l, scaled root mean squared error, $srmse_l$, and correlation, $corr_l$, for $l = 1, \ldots, 44$ CMIP5 models used in this study. Models with compatibilities greater than 0.001 are italicized.

l	*Model*	p_l	$srmse_l$	$corr_l$	l	*Model*	p_l	$srmse_l$	$corr_l$
1	ACCESS1-0	< 0.001	0.351	0.598	23	GFDL-ESM2M	< 0.001	0.276	0.611
2	ACCESS1-3	< 0.001	0.411	0.659	24	GISS-E2-H p1	< 0.001	0.359	0.751
3	BCC-CSM-1	< 0.001	0.241	0.764	25	GISS-E2-H-CC p1	< 0.001	0.182	0.73
4	BCC-CSM-1-M	< 0.001	0.208	0.711	26	GISS-E2-R p1	< 0.001	0.47	0.746
5	*BNU-ESM*	*0.010*	0.000	0.705	27	GISS-E2-R-CC p1	< 0.001	0.415	0.708
6	*CanESM2*	*0.015*	0.427	0.729	28	HadGEM2-AO	< 0.001	0.341	0.672
7	*CCSM4*	*0.199*	0.214	0.762	29	*HadGEM2-CC*	*0.137*	0.165	0.43
8	*CESM1-BGC*	*0.241*	0.166	0.709	30	HadGEM2-ES	0.000	0.366	0.667
9	CESM1-CAM5	< 0.001	0.527	0.764	31	INMCM4	< 0.001	0.336	0.658
10	CESM1-CAM5-1-FV2	< 0.001	0.482	0.711	32	IPSL-CM5A-LR	< 0.001	0.305	0.772
11	*CESM1-FASTCHEM*	*0.089*	0.105	0.752	33	IPSL-CM5A-MR	< 0.001	0.302	0.75
12	*CESM1-WACCM*	*0.078*	0.068	0.692	34	IPSL-CM5B-LR	< 0.001	0.16	0.658
13	CMCC-CESM	< 0.001	0.169	0.36	35	MIROC-ESM	< 0.001	0.442	0.729
14	CMCC-CM	< 0.001	0.39	0.63	36	MIROC-ESM-CHEM	< 0.001	0.377	0.688
15	CMCC-CMS	< 0.001	0.211	0.46	37	MIROC5	< 0.001	0.474	0.714
16	CNRM-CM5	< 0.001	0.514	0.765	38	MPI-ESM-LR	< 0.001	0.225	0.728
17	CSIRO-Mk3-6-0	< 0.001	0.45	0.706	39	*MPI-ESM-MR*	*0.001*	0.32	0.749
18	EC-EARTH	< 0.001	0.287	0.758	40	MPI-ESM-P	< 0.001	0.245	0.734
19	FGOALS-g2	< 0.001	0.487	0.723	41	MRI-CGM3	< 0.001	0.428	0.624
20	FIO-ESM	< 0.001	0.303	0.712	42	MRI-ESM1	< 0.001	0.369	0.624
21	*GFDL-CM3*	*0.200*	0.264	0.628	43	NorESM1-M	< 0.001	0.484	0.715
22	GFDL-ESM2G	< 0.001	0.446	0.691	44	*NorESM1-ME*	*0.001*	0.397	0.658

LIST OF FIGURES

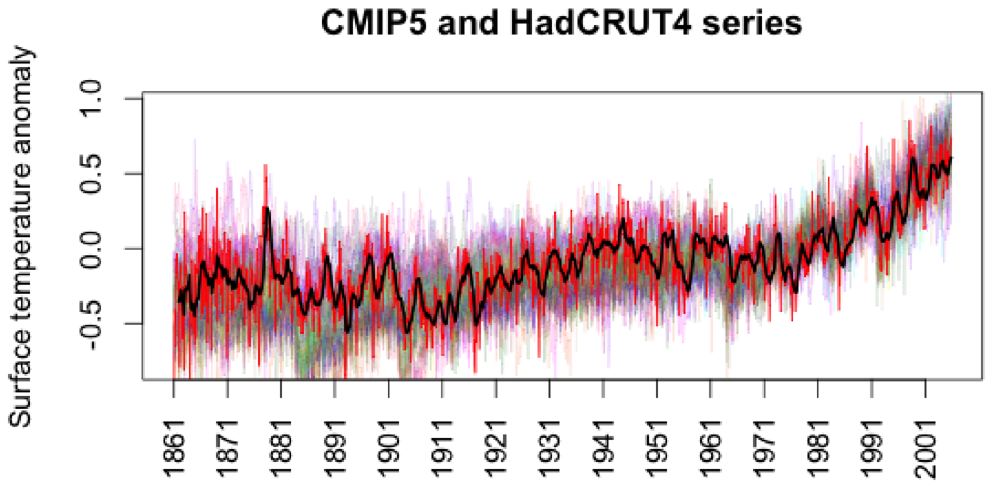
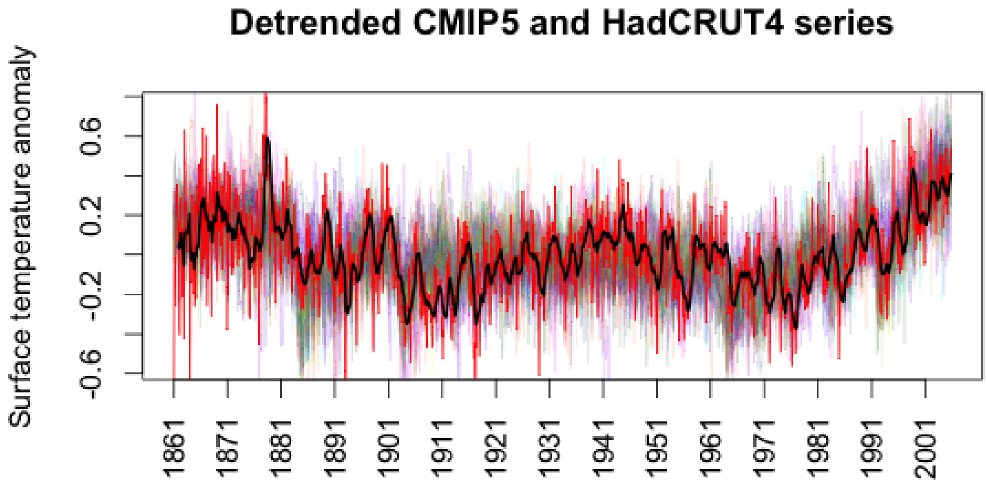


FIG. 1. Anomaly time sequence plots for 44 CMIP5 outputs of monthly global average near-surface air temperature anomalies (pastels), and the HadCRUT4 observational sequence (red), 1861–2005. The black line is a 12-month running mean computed from the HadCRUT4 (red line) data.



802 FIG. 2. Linearly detrended anomaly time sequence plots for 44 CMIP5 outputs of monthly global average near-surface air temperature anomalies (pastels), and the HadCRUT4 observational sequence (red), 1861–2005. The black line is a 12-month running mean computed from the HadCRUT4 (red line) detrended data.

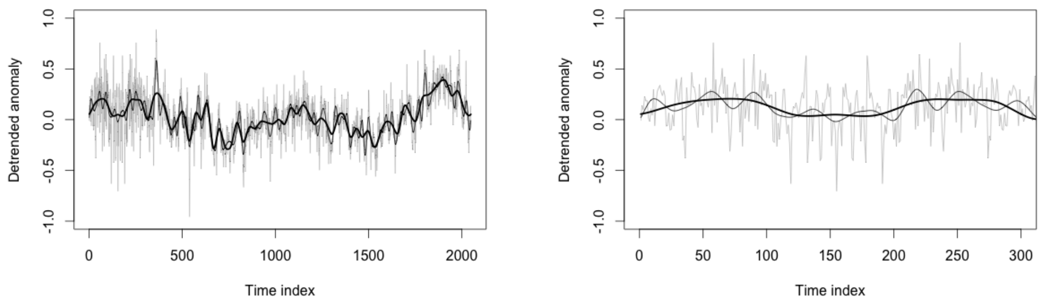


FIG. 3. Detrended, padded time sequence for the HadCRUT4 observational sequence (light gray). The reconstructed sequence using wavelet levels 1 through $\check{J} = 6$ is shown by the thin black line. The reconstructed sequence using wavelet levels up to and including level $\check{J} = 5$ is shown by the thick black line. Left panel: the entire sequence. Right panel: only the first 300 time points.

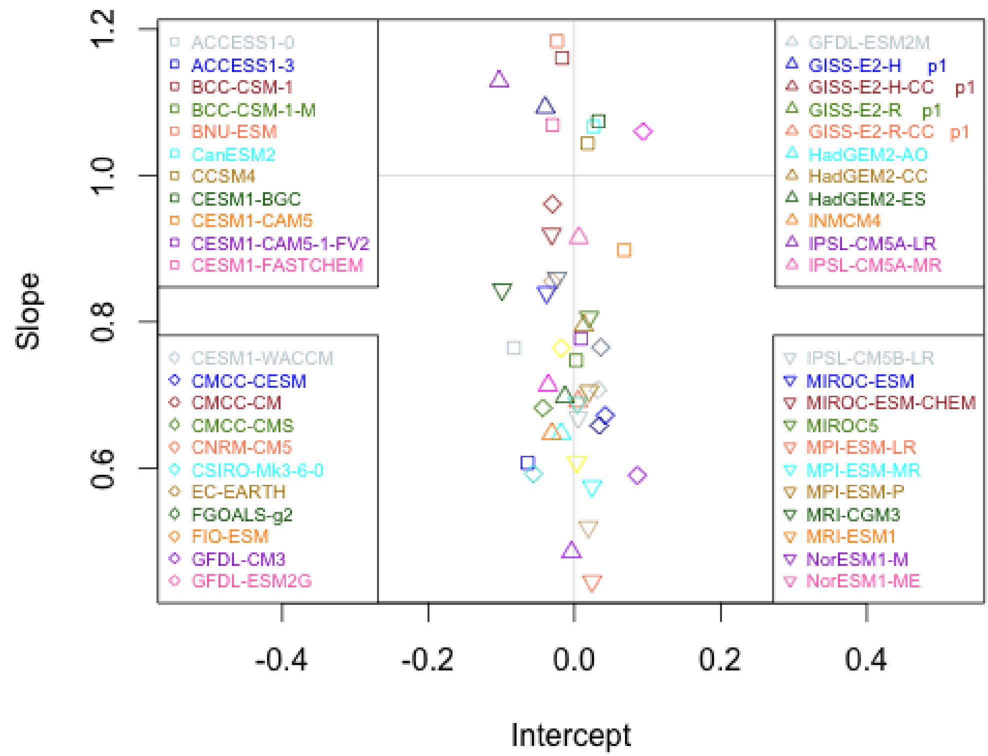


FIG. 4. Plot of $\hat{\beta}_l$ versus $\hat{\alpha}_l$, for $l = 1, \ldots, 44$ CMIP5 models. Symbols and colors vary in order to differentiate visually among models. The filled circle at plot coordinate (0,1) represents perfect agreement between model output and observations.

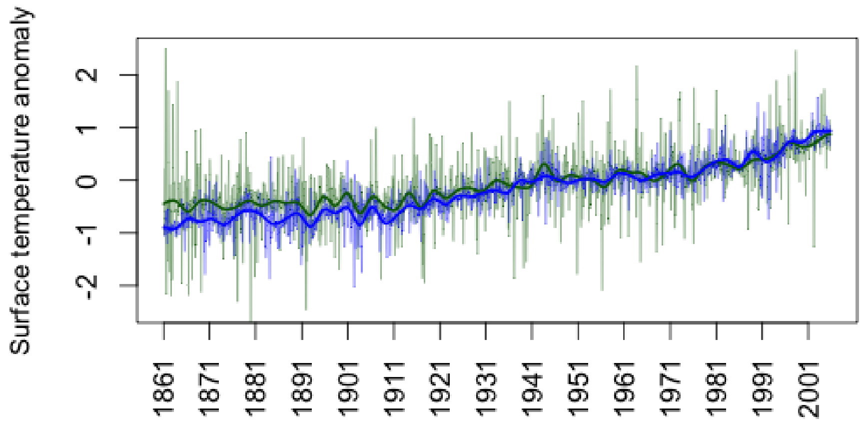


FIG. 5. Plots of one pair of resampled time sequences obtained from the CCSM4 model (blue) and the HadCRUT4 observations (green). Smoothed versions using wavelet decomposition levels up to and including $\check{J} = 5$ are shown by the thick blue and green lines.

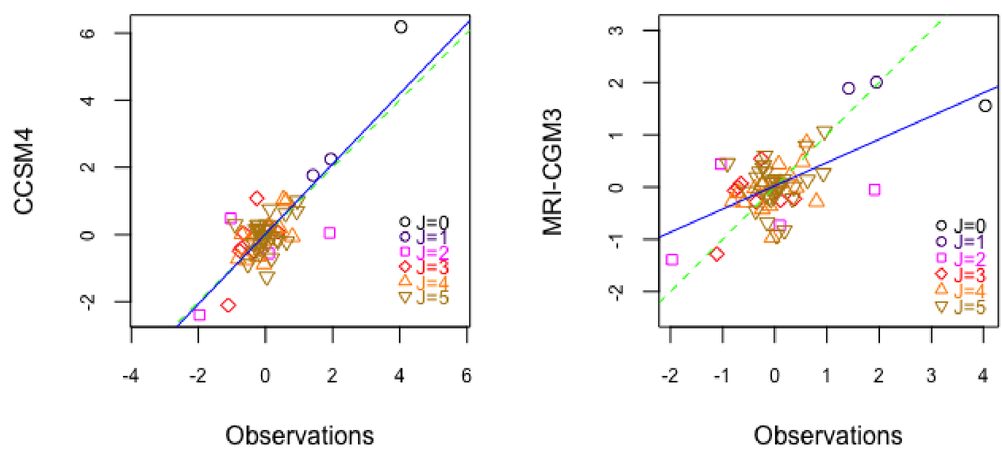


FIG. 6. Plots of climate-scale wavelet coefficients of two climate model-output time sequences on those of the HadCRUT4 observational time sequence. The two models are CCSM4 (left panel) and MRI-CGM3 (right panel). Each point in the plot is color- and symbol-coded to show the level of the wavelet decomposition to which it belongs. The solid blue lines are the regression lines determined by the fit to the scatterplots, and the dashed green lines are 45° lines.

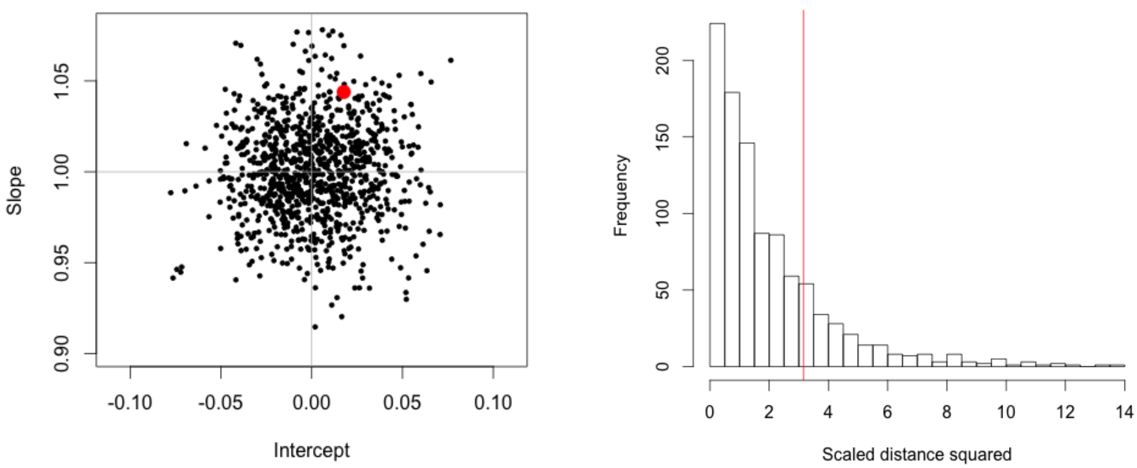


FIG. 7. Left: Scatterplot of $\left\{(\hat{\alpha}^*_{bl}, \hat{\beta}^*_{bl}) : \ b = 1, \ldots, 1000\right\}$ (black dots). The value of $(\hat{\alpha}_l, \hat{\beta}_l)$ for the CCSM4 model is given by the large red dot. Right: Histogram of $\left\{Q^*_{bl} : \ b = 1, \ldots, 1000\right\}$ for l given by the CCSM4 model. The actual Q_l for the CCSM4 model is located at the red vertical line.

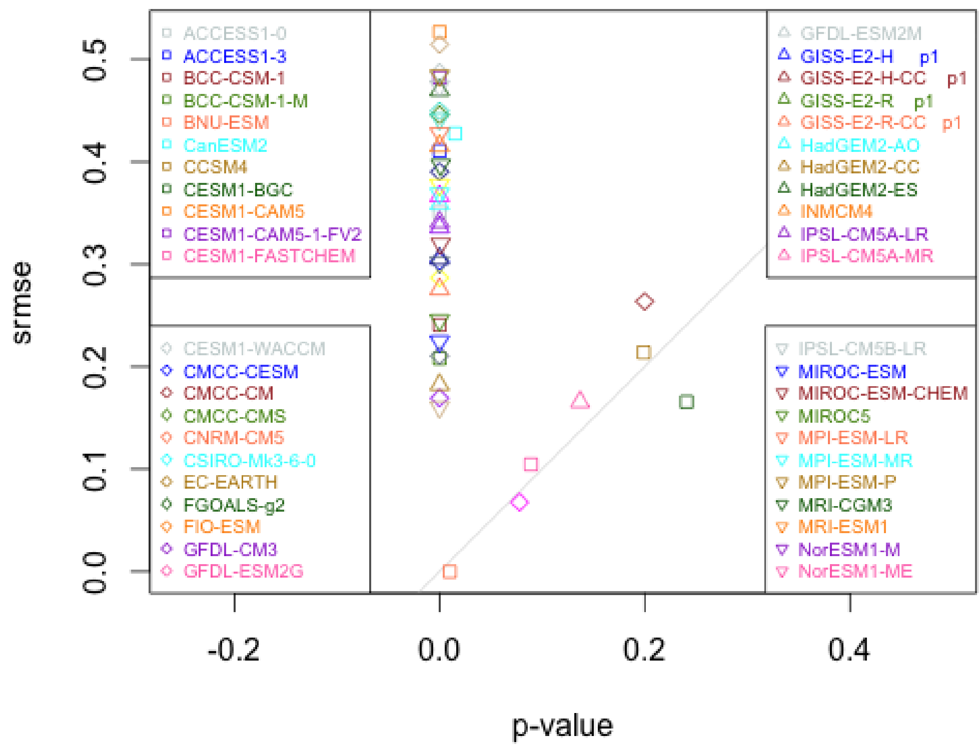
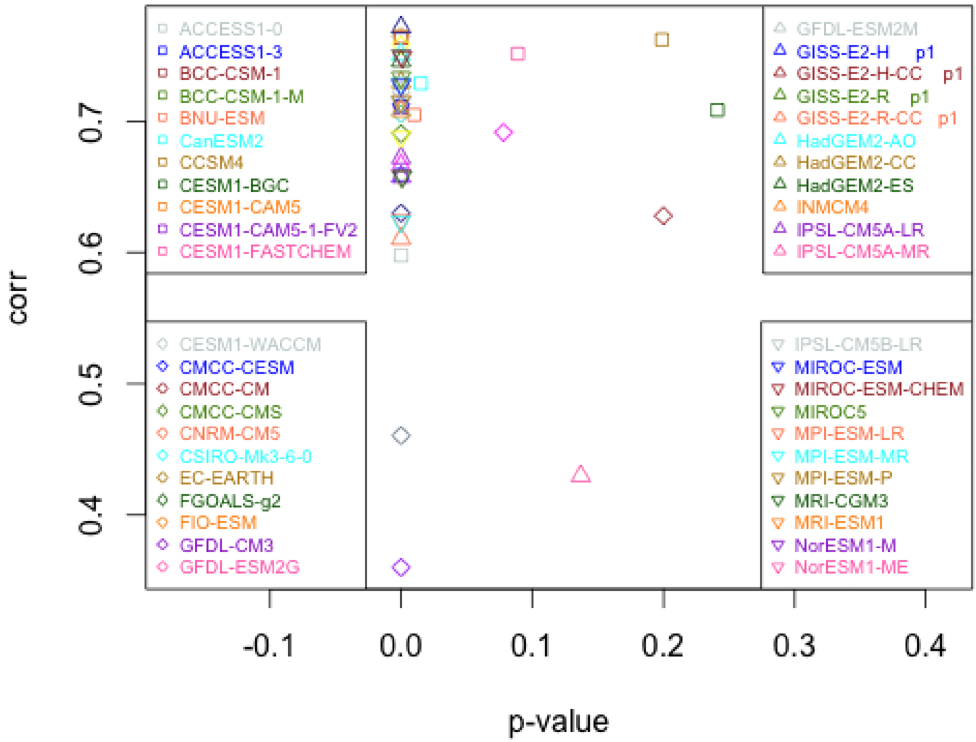


FIG. 8. Scatterplot of $srmse_l$ versus p_l; values are given in Table 3. The 45° line is shown in gray. Symbols and colors vary in order to differentiate visually among models.



825 FIG. 9. Scatterplot of $corr_l$ versus p_l; values are given in Table 3. The o45° line is shown in gray. Symbols
826 and colors vary in order to differentiate visually among models.

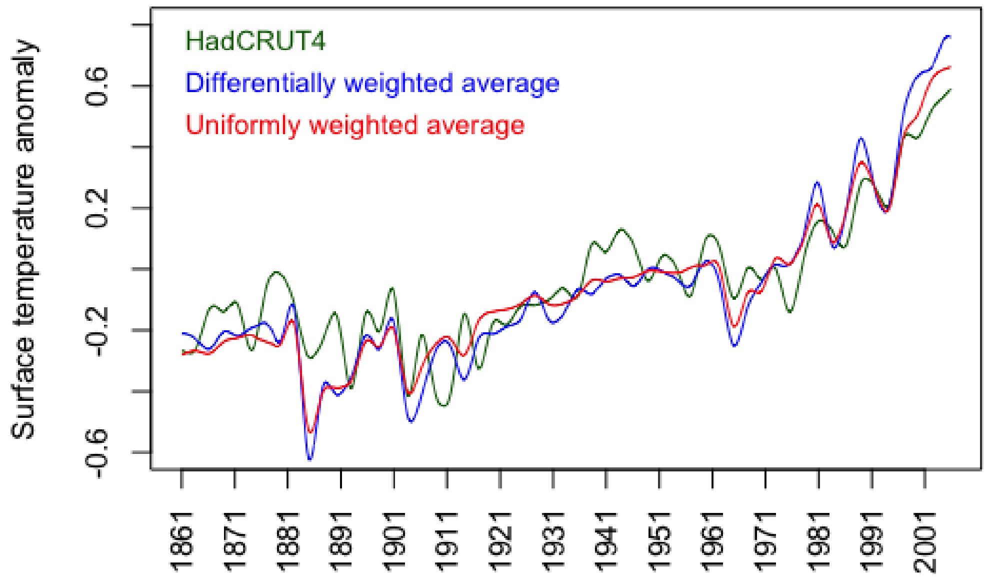
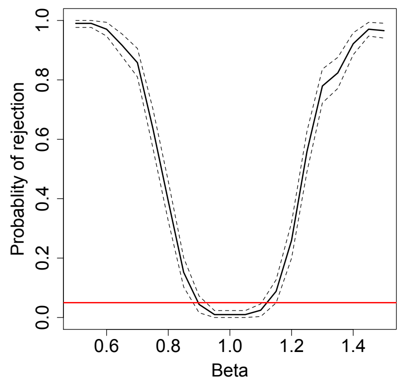
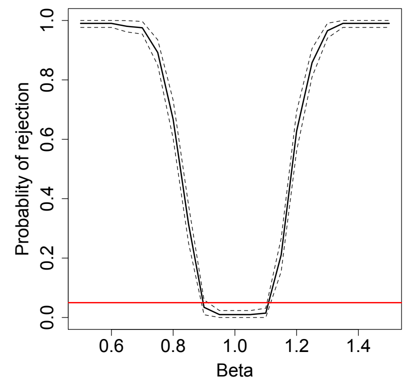


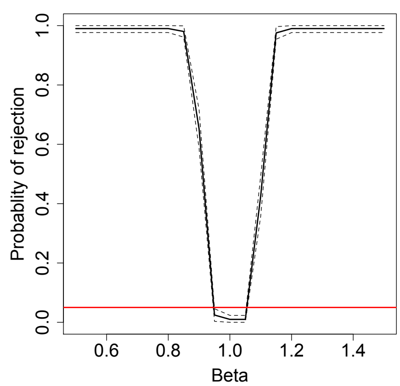
FIG. 10. Time sequences of HadCRUT4 observations (green), and the differentially and uniformly weighted multi-model averages given by Eq. (44) (blue and red, respectively), reconstructed using wavelet decomposition levels up to and including $\check{J} = 5$, and with trend added back in.


Probablity of rejection
0.0
0.2
0.4
0.6
0.8
1.0
0.6
0.8
1.0
1.2
1.4
Beta

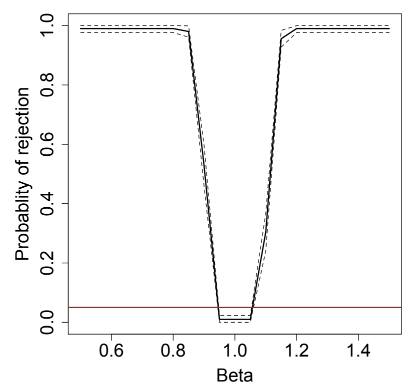
(a)


Probablity of rejection
0.0
0.2
0.4
0.6
0.8
1.0
0.6
0.8
1.0
1.2
1.4
Beta

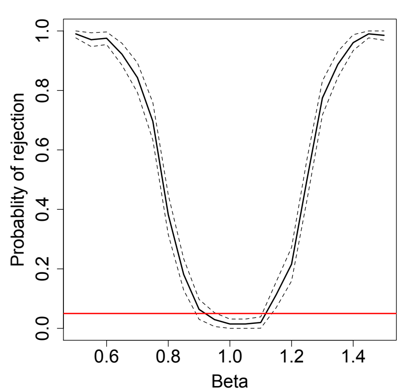
(b)


Probablity of rejection
0.0
0.2
0.4
0.6
0.8
1.0
0.6
0.8
1.0
1.2
1.4
Beta

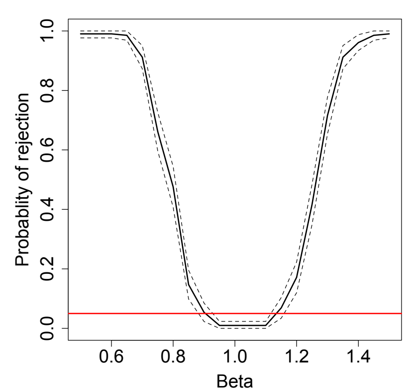
(c)


Probablity of rejection
0.0
0.2
0.4
0.6
0.8
1.0
0.6
0.8
1.0
1.2
1.4
Beta

(d)


Probablity of rejection
0.0
0.2
0.4
0.6
0.8
1.0
0.6
0.8
1.0
1.2
1.4
Beta

(e)


Probablity of rejection
0.0
0.2
0.4
0.6
0.8
1.0
0.6
0.8
1.0
1.2
1.4
Beta

(f)

Fig. B1. Power curves from different simulation scenarios. Panel (a) shows the power curve corresponding to sample size $N = 600$, noise variance $V = 0.01$, and $\check{J} = 3$ coarse wavelet levels. Panel (b) uses $N = 1000$, $V = 0.01$, $\check{J} = 3$. Panel (c) uses $N = 600$, $V = 0.01$, $\check{J} = 5$. Panel (d) uses $N = 1000$, $V = 0.01$, $\check{J} = 5$. Panel (e) uses $N = 600$, $V = 0.2$, $\check{J} = 5$. Panel (f) uses $N = 1000$, $V = 0.2$, $\check{J} = 5$. The dashed lines are point-wise 95% confidence intervals for the power function.