Probabilistic Evaluation of Competing Climate Models

Amy Braverman[1], Snigdhansu Chatterjee[2], Megan Heyman[3], and Noel Cressie[4]

[1]Jet Propulsion Laboratory, California Institute of Technology, Mail Stop 158-242, 4800 Oak Grove Drive, Pasadena, CA 91109
[2]University of Minnesota, 313 Ford Hall, 224 Church St. S.E., Minneapolis, MN 55455
[3]Rose-Hulman Institute of Technology, G-205 Crapo Hall, 5000 Wabash Ave., Terre Haute, IN 47803
[4]University of Wollongong, Northfields Ave., Wollongong, NSW 2522, Australia

Correspondence to: Amy.Braverman (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 climate system. A standard paradigm for assessing the quality of climate model simulations is to compare what these models produce for past and present time periods, to observations of the past and present. Many of these comparisons are based on simple summary statistics called metrics. In this article, we propose an alternative: evaluation of competing climate models through probabilities derived from tests of the hypothesis that climate-model-simulated and observed time sequences share common climate-scale signals. The probabilities are based on the behavior of summary statistics of climate model output and observational data, over ensembles of pseudo-realizations. These are obtained by partitioning the original time sequences into signal and noise components, and using a parametric bootstrap to create pseudo-realizations of the noise sequences. The statistics we choose come from working in the space of decorrelated and dimension-reduced wavelet space. Here we compare monthly sequences of CMIP5 model output of average global near-surface temperature anomalies to similar sequences obtained from the well-known HadCRUT4 data set, as an illustration.

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 the 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,

both across models and between runs of the same model using different initial conditions and parameter settings. To organize the effort the Coupled Model Intercomparison Project (CMIP) was established "to provide climate scientists with a database of coupled GCM simulations under standardized boundary conditions," and "to attempt to discover why different models give different output in response to the same input, or (more typically) to simply identify aspects of the simulations in which 'consensus' in model predictions or common problematic features exist" (Covey et al., 2003). CMIP, now beginning its sixth incarnation (CMIP6), has grown to facilitate the use of multimodel (Tebaldi and Knutti, 2007) and perturbed physics (Murphy et al., 2004; Deser et al., 2010) ensembles as a means of quantifying uncertainties in future projections of climate change.

An enormous literature exists on the use of climate models, and ensembles of model outputs, to make predictions of future climate conditions and quantify reliabilities of those predictions. A basic strategy for quantifying reliability of individual model runs is to assess their performance, over the past and present, against observations. Baumberger et al. (2017) call the ability of climate models to generate simulations that agree with observed data, "empirical accuracy". The supposition is that agreement of climate model simulations with observations is an indication that the physics of the climate model are correct. Assuming that the physics of the future is the same as the physics of today, this implies that future projections of models that achieve empirical accuracy are more reliable than the projections of those models that do not. There are many reasons to believe that things aren't that simple (Baumberger et al., 2017; Sanderson and Knutti, 2012), but nonetheless there are plenty of examples of the use of observations to determine how members of model generated ensembles of predictions should be weighted (Annan and Hargeaves, 2010; Boe and Terray, 2015; Hung et al., 2013; Giorgi and Mearns, 2002; Suh and Oh, 2012).

Even if empirical accuracy is not sufficient to establish reliability of future projections, there are other reasons why one might want to compare climate model simulations to observations. First, there is diagnostic value in understanding the ways in which climate model simulations agree or disagree with observed conditions (Kiehl, 2006; Watanabe et al., 2010; Meehl et al., 2009). Second, there is growing consensus that CMIP activities should include systematic evaluation of models against observations to document improvements in the models over time and identify those aspects of model performance most in need of improvement (Eyring et al., 2016). The World Climate Research Program (WCRP) Working Group on Numerical Experiments (WGNE) has established a Diagnostics and Metrics Panel to oversee the development of "metrics" that can be used for these purposes. Metrics endorsed by the Panel at present tend to be simple descriptive summary statistics such as root mean squared error (RMSE) over a time series or spatial field (see Gleckler et al. (2008), for example).

Descriptive metrics are valuable as relative measures of the goodness-of-fit of climate model simulations to observations. One can say that the RMSE, against observations, of one model run is lower than that of another. However, it's hard to know how to interpret metric values in an absolute sense: how does the value of the metric relate to the probability that a model is "right" in its representation of an observed physical process? That question is malformed until we are precise about what "right" means. We must articulate a specific hypothesis about the relationship between observed and climate-model-simulated data; the model is deemed to be "right" if a formal statistical test of that hypothesis is not rejected, at an agreed upon level of significance. The p-value of this test can be interpreted as a measure of the compatibility of the data with the hypothesis (Wasserstein and Lazar, 2016). This compatibility measure can be used as a probability-scale metric of the degree to which the model simulation is a "right" representation of the observed data.

In this article, we present the statistical machinery for deriving compatibility measures between climate model-simulated and observed time sequences. The null hypothesis we test is that the coarse-time-scale coefficients of wavelet decompositions of the two sequences, are the same. This allows for the possibility that, in the time domain, the sequences do not match exactly, but rather share longer-term, climate-scale behavior. Specifically, we break the time sequences of observations and climate model-generated output into *two* components: low-frequency sequences described by coarse-level wavelet coefficients, and high-frequency (possibly non-stationary) sequences described by an integrated autoregressive moving average (ARIMA) model. The coarse-level wavelet coefficients characterize decadal and multi-decadal-scale oscillatory patterns, which we call "climate signal", while the ARIMA processes characterize temporal dependence at finer time scales and which we call "climate noise". Our measure of similarity is the squared euclidean distance between vectors of climate signal wavelet coefficients. The high-frequency, climate noise might be interpreted as "weather", and *do not* contribute to this measure of similarity. To generate sampling distributions under the null hypothesis, we employ a parametric bootstrap in the time domain, based on the ARIMA models fit to the climate noise. We demonstrate our method by computing the compatibilities of 139 CMIP5 historical model runs, of 44 different models, simulating monthly global near-surface temperature anomalies. We use the HadCRUT4 monthly global near-surface temperature anomaly data set as our observational benchmark.

The remainder of this paper is organized as follows. Section 2 describes the statistical model that relates model-generated output, observations, and true climate to one another. Section 3 defines the hypothesis testing framework that is crucial to our evaluation, along with the algorithm we use to implement it. In Section 4 we demonstrate our method and algorithm by evaluating the output of CMIP5 climate models against observations. Conclusions follow in Section 5.

2 A wavelet-based statistical model for true climate, model-generated, and observed time sequences

Consider a single climate variable (e.g., global average near-surface temperature) whose true value is generically denoted as Y. Define $\boldsymbol{Y} \equiv (Y_1, \ldots, Y_t, \ldots, Y_T)'$ to be a column vector of length T representing a sequence of values of Y through time up to the present. Observations are represented by the T-dimensional column vector $\boldsymbol{Z}$, and the l-th climate model's simulated time sequence is denoted by $\boldsymbol{X}_l$, $l = 1, 2, \ldots, L$ where L is the number of model runs.

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

$$\boldsymbol{X}_l = \boldsymbol{Y} + \boldsymbol{e}_l \qquad \text{and} \qquad \boldsymbol{Z} = \boldsymbol{Y} + \boldsymbol{e}_0, \tag{1}$$

where $\boldsymbol{e}_l$ is the error of the l-th climate model sequence, and $\boldsymbol{e}_0$ is the error on the observations (Rougier, 2007). This is the standard "truth-plus-error" statistical model often discussed in the climate literature (Annan and Hargeaves, 2010).

Direct comparison of $\boldsymbol{X}_l$ to $\boldsymbol{Z}$, say by computing $D_l = ||\boldsymbol{X}_l - \boldsymbol{Z}||^2$ (or a weighted version), suffers from several problems that make the result difficult to interpret. First, $\boldsymbol{X}_l$ and $\boldsymbol{Z}$ are not expected to match element-by-element. We would like to capture some notion of common structure, rather than pointwise agreement in time. Second, all these time sequences will exhibit temporal dependence, so any methodology and its associated theory needs to account for it. Both issues are effectively addressed by transforming the time sequences using a wavelet decomposition.

The wavelet decomposition is a decorrelator, just like the usual Fourier spectral decomposition, but wavelets capture local behavior through functions that are of compact support, multi-resolutional, and translational within a resolution. Lin and Franzke (2015) have showed that wavelets can capture multiresolution temporal structure in global average near-surface temperatures.

In wavelet analysis, the Discrete Wavelet Transform (DWT) is

$$\mathcal{C}_{\boldsymbol{X}} \equiv W\boldsymbol{X}, \tag{2}$$

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* $\mathcal{C}_{\boldsymbol{X}}$ (Percival and Walden, 2006). The choice of wavelet basis functions (father and mother wavelets) will determine the form of W.

We augment the model given in Eq. (1) as follows. Let $\boldsymbol{Y}^s$ and $\boldsymbol{Y}^n$ denote the climate "signal" and "noise" components of $\boldsymbol{Y}$, where climate-signal is *defined* by the number of coarse-scale wavelet decomposition levels that distinguish climate-scale variability from weather-scale variability. This partitioning depends on the scientific problem being addressed, the hypothesis of interest, and the assumptions the analyst is willing to make. Define $\boldsymbol{X}_l^s$, $\boldsymbol{X}_l^n$, $\boldsymbol{Z}^s$, and $\boldsymbol{Z}^n$ analogously. Then, since the wavelet transformation is linear,

$$\boldsymbol{Y} = \boldsymbol{Y}^s + \boldsymbol{Y}^n, \qquad \boldsymbol{X}_l = \boldsymbol{Y}^s + \boldsymbol{Y}^n + \boldsymbol{e}_l, \qquad \boldsymbol{Z} = \boldsymbol{Y}^s + \boldsymbol{Y}^n + \boldsymbol{e}_0. \tag{3}$$

$$\mathcal{C}_{\boldsymbol{Y}} = \mathcal{C}_{\boldsymbol{Y}^s} + \mathcal{C}_{\boldsymbol{Y}^n}, \qquad \mathcal{C}_{\boldsymbol{X}_l} = \mathcal{C}_{\boldsymbol{Y}^s} + (\mathcal{C}_{\boldsymbol{Y}^n} + \mathcal{C}_{\boldsymbol{e}_l}), \qquad \mathcal{C}_{\boldsymbol{Z}} = \mathcal{C}_{\boldsymbol{Y}^s} + (\mathcal{C}_{\boldsymbol{Y}^n} + \mathcal{C}_{\boldsymbol{e}_0}). \tag{4}$$

The terms in parentheses in Eq. (4) cannot be separately identified, so they are combined and we consider them to be residual errors.

The key assumption that we make is that $\mathcal{C}_{\boldsymbol{Z}}$ can be denoised, at least in an asymptotic sense, to leave behind only the wavelet coefficients associated with climate-signal, $\mathcal{C}_{\boldsymbol{Y}^s}$. Suppose that T, the length of the time sequences, is an exact power of two. If it is not, the sequences can be padded appropriately as discussed below in Section 3. Let $\check{J}$ be a constant, $\check{J} \leq J = log_2 T$, that specifies the number of coarse-scale wavelet decomposition levels that define climate-signal in the wavelet-level hierarchy. Let $\mathcal{S}(\mathcal{C}_{\boldsymbol{X}}, \check{J})$ be a smoothing function that operates on a generic vector of wavelet coefficients, $\mathcal{C}_{\boldsymbol{X}}$, by setting elements corresponding to levels above the first $\check{J}$, to zero. Let $\mathcal{T}(\mathcal{C}_{\boldsymbol{X}}, \check{J})$ be a truncation function that deletes the trailing zero elements in $\mathcal{C}_{\boldsymbol{X}}$. Then,

$$\begin{aligned}
\mathcal{C}_{\boldsymbol{X}} &= \left(\gamma_{00}, \gamma_{01}, \ldots, \gamma_{(\check{J}-1)2^{(\check{J}-1)}}, \gamma_{\check{j}1}, \ldots, \gamma_{(J-1)(2^{J-1})}\right)', \\
\mathcal{S}(\mathcal{C}_{\boldsymbol{X}}, \check{J}) &= \left(\gamma_{00}, \gamma_{01}, \ldots, \gamma_{(\check{J}-1)2^{(\check{J}-1)}}, 0, \ldots\ldots\ldots, 0\right)', \\
\boldsymbol{\Gamma}_{\boldsymbol{X}} \equiv \mathcal{T}\left(\mathcal{S}(\mathcal{C}_{\boldsymbol{X}}, \check{J}), \check{J}\right) &= \left(\gamma_{00}, \gamma_{01}, \ldots, \gamma_{(\check{J}-1)2^{(\check{J}-1)}}\right)',
\end{aligned} \tag{5}$$

where γ_{jk} is the k-th wavelet coefficient at level j. Our assumption is that $\boldsymbol{\Gamma}_{\boldsymbol{Z}} = \boldsymbol{\Gamma}_{\boldsymbol{Y}^s}$, that the wavelet coefficients that define the true climate-signal can be recovered from the observations. Of course, this requires us to specify an appropriate value of $\check{J}$. As noted above, this choice will be problem-dependent. The corresponding smoothed time sequence is $S(\boldsymbol{X}, \check{J}) = W'\mathcal{S}(\mathcal{C}_{\boldsymbol{X}}, \check{J})$.

We now establish some important notation for further specifying the statistical models. Write $\boldsymbol{X}_l = (X_l(1), \ldots, X_l(T))'$, for $l = 1, \ldots, L$, and $\boldsymbol{Z} = (Z(1), \ldots, Z(T))'$. We model $X_l(t)$ and $Z(t)$ as follows:

$$X_l(t) = \alpha_l + \beta_l t + \gamma_{l2} V_l(t/T) + \mu_l(t) + e_l(t), \text{ for } t = 1, \ldots, T,\ l = 1, \ldots, L, \tag{6}$$

$$Z(t) = \alpha_0 + \beta_0 t + \gamma_{02} V_0(t/T) + \mu_0(t) + e_0(t), \text{ for } t = 1, \ldots, T, \tag{7}$$

where α_l and β_l are linear trend coefficients, $V_l(\cdot)$ and γ_{l2} are scaling functions and coefficients respectively, $l = 0, \ldots, L$. Note that the case $l = 0$ refers to quantities in the statistical model of the observations. In Eqs. (6) and (7),

$$\mu_l(t) = \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} \gamma_{ljk} W_{jk}(t/T), \text{ for } l = 0, \ldots, L,\ t = 1, \ldots, T, \tag{8}$$

where $W_{jk}(\cdot)$ is a fixed family of wavelet basis functions. The vectors of climate-scale wavelet coefficients are

$$\boldsymbol{\Gamma}_{\boldsymbol{X}_l} = \left(\gamma_{l00}, \ldots, \gamma_{l(J-1)(2^{J-1})}\right)', \text{ for } l = 1, \ldots, L, \tag{9}$$

and

$$\boldsymbol{\Gamma}_{\boldsymbol{Z}} = \left(\gamma_{000}, \ldots, \gamma_{0(J-1)(2^{J-1})}\right)'. \tag{10}$$

Further, we assume that the noise terms, $e_l(t)$ and $e_0(t)$, are all mutually independent with means equal to zero but potentially unequal variances, for $l = 1, \ldots, L$; $\mathrm{E}\left(e_l^2(t)\right) = \sigma_l^2(t)$, and $\mathrm{E}\left(e_0^2(t)\right) = \sigma_0^2(t)$.

We will apply the same wavelet transform to detrended versions of $\{\boldsymbol{X}_l\}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$, and work in the equivalent space of wavelet coefficients. Thus, we can now clearly define what we mean by common structure of two time sequences: they share the same climate-scale wavelet coefficients. We think of this as a null hypothesis,

$$H_0 : \boldsymbol{\Gamma}_{\boldsymbol{X}_l} = \boldsymbol{\Gamma}_{\boldsymbol{Z}}. \tag{11}$$

3 Hypothesis testing framework

To carry out a test of the hypothesis H_0 in Eq. (11), we must identify a test statistic and generate the distribution of that statistic under the assumption of H_0.

3.1 Test statistic that captures a relationship to the true climate

The test statistics that we use are based on a weighted squared distance between the climate-scale wavelet coefficients of $\boldsymbol{X}_l$ and $\boldsymbol{Z}$. Recall,

$$\boldsymbol{\Gamma}_{\boldsymbol{Z}} = \mathcal{T}\left(\mathcal{S}(\mathcal{C}_{\boldsymbol{Z}}, \check{J}), \check{J}\right) = \left(\gamma_{000}, \ldots, \gamma_{0(\check{J}-1)(2^{\check{J}-1})}\right)', \tag{12}$$

$$\boldsymbol{\Gamma}_{\boldsymbol{X}_l} = \mathcal{T}\left(\mathcal{S}(\mathcal{C}_{\boldsymbol{X}_l}, \check{J}), \check{J}\right) = \left(\gamma_{l00}, \ldots, \gamma_{l(\check{J}-1)(2^{\check{J}-1})}\right)', \text{ for } l = 1, \ldots, L, \tag{13}$$

for a fixed value of $\check{J}$. These vectors are of length $\lambda = \sum_{j=0}^{\check{J}-1} \sum_{k=0}^{2^j-1} 1$, which is the total number of wavelet coefficients corresponding to the climate-signal. We define the test statistic D_l,

$$D_l \equiv \left(\hat{\mathbf{\Gamma}}_{\boldsymbol{X}_l} - \hat{\mathbf{\Gamma}}_{\boldsymbol{Z}}\right)' \mathbf{\Omega} \left(\hat{\mathbf{\Gamma}}_{\boldsymbol{X}_l} - \hat{\mathbf{\Gamma}}_{\boldsymbol{Z}}\right), \quad \mathbf{\Omega} = \text{diag}(\omega_{11}, \omega_{22}, \ldots, \omega_{\lambda\lambda}), \tag{14}$$

where $\hat{\mathbf{\Gamma}}_{\boldsymbol{X}_l}$ and $\hat{\mathbf{\Gamma}}_{\boldsymbol{Z}}$ are estimates of $\mathbf{\Gamma}_{\boldsymbol{X}_l}$ and $\mathbf{\Gamma}_{\boldsymbol{Z}}$ computed from $\boldsymbol{X}_l$ and $\boldsymbol{Z}$, respectively, and $\mathbf{\Omega}$ is an $\lambda \times \lambda$ diagonal matrix of weights in which the diagonal element corresponding to γ_{ljk} is proportional to $T/2^j$, for $k = 0, 1, \ldots, 2^j - 1$, and $l = 0, 1, \ldots, L$. This makes the weights proportional to the number of time points influenced by the wavelet coefficients. We rescale these diagonal entries so that they sum to one in order to facilitate easier interpretation as weights.

3.2 Simulating the null distribution of the test statistic

In what follows, it is crucial to obtain good estimates of the test statistic's variance under $H_0 : \mathbf{\Gamma}_{\boldsymbol{X}_l} = \mathbf{\Gamma}_{\boldsymbol{Z}}$ against the alternative $H_A : \mathbf{\Gamma}_{\boldsymbol{X}_l} \neq \mathbf{\Gamma}_{\boldsymbol{Z}}$; $l = 1, \ldots, L$. We obtain variance estimates by generating B "pseudo-realizations" of a time sequence from a single parent time sequence, under H_0. Then, for each pseudo-realization indexed by b, we detrend, perform the wavelet decomposition, and compute the test statistic to obtain B resampled values of D_l, $\{D_{lb}^* : b = 1, \ldots, B\}$. The empirical variance of this sample is an approximation to the sampling variance of D_l under H_0.

Starting with the original sequences, $\boldsymbol{Z}$ of length N_0 and $\boldsymbol{X}_l$ of length N_l, we perform the following steps.

1. Set B (the number of trials).
2. Obtain $\tilde{\tilde{\boldsymbol{X}}}_l$ and $\tilde{\tilde{\boldsymbol{Z}}}$ as follows:
   (a) Perform simple linear regression of $\boldsymbol{Z}$ on the sequence $\{t : t = 1, 2, \ldots, N_0\}$ to obtain the regression intercept and slope, $\hat{\alpha}_0$ and $\hat{\beta}_0$.
   (b) Perform simple linear regression of $\boldsymbol{X}_l$ on the sequence $\{t : t = 1, 2, \ldots, N_l\}$ to obtain the regression intercept and slope, $\hat{\alpha}_l$ and $\hat{\beta}_l$.
   (c) Set $\tilde{\tilde{Z}}(t) = Z(t) - \hat{\alpha}_0 - \hat{\beta}_0 t$, for $t = 1, 2, \ldots, N_0$.
   (d) Set $\tilde{\tilde{X}}_l(t) = X_l(t) - \hat{\alpha}_l - \hat{\beta}_l t$, for $t = 1, 2, \ldots, N_l$.

   Retain the computed values of the trend coefficients, $\left(\hat{\alpha}_0, \hat{\beta}_0\right)$ and $\left(\hat{\alpha}_l, \hat{\beta}_l\right)$.
3. If either N_0 or N_l is not an exact power of two, then pad $\tilde{\tilde{\boldsymbol{Z}}}$ and $\tilde{\tilde{\boldsymbol{X}}}_l$ so that both their lengths are equal to $T = 2^{\lceil \log_2 N \rceil}$, $N = \max(N_0, N_l)$, where $\lceil \cdot \rceil$ is the ceiling function. This padding is implemented by reflection at the beginning and end of the sequences. Call the padded sequences, $\tilde{\boldsymbol{Z}}$ and $\tilde{\boldsymbol{X}}_l$. If no padding is required, set $\tilde{\boldsymbol{Z}} = \tilde{\tilde{\boldsymbol{Z}}}$ and $\tilde{\boldsymbol{X}}_l = \tilde{\tilde{\boldsymbol{X}}}_l$. Note that padding sequences in this way is standard practice in wavelet-based analysis (Ogden, 1997). In our application below, we actually do not need to implement this step since our time sequences are of lengths that are exact powers of two.
4. Set $J = \log_2 T$ and $\check{J}$ equal to the number of levels in the wavelet decomposition that constitute climate-signal.

5. Perform the J-level wavelet decomposition on $\tilde{\boldsymbol{Z}}$ to obtain the set of climate-signal wavelet coefficients $\hat{\boldsymbol{\Gamma}}_{\boldsymbol{Z}}$ $= \left(\hat{\gamma}_{000}, \hat{\gamma}_{001}, \ldots, \hat{\gamma}_{0(\check{J}-1)2^{(\check{J}-1)}}\right)$. Our choice of wavelet basis ensures that the coefficient of the scaling functions, γ_{l2} and γ_{02} in Eqns. (6) and (7) are can be assumed to be zero, because of the linear regression implemented in Step 2. Consequently we do not include these terms from the climate-signal's wavelet-coefficient vectors.

6. Compute $\hat{\boldsymbol{\mu}}_0 = (\hat{\mu}_0(1), \hat{\mu}_0(2), \ldots, \hat{\mu}_0(T))'$ from $\hat{\boldsymbol{\Gamma}}_{\boldsymbol{Z}}$:

$$\hat{\mu}_0(t) = \sum_{j=0}^{\check{J}-1} \sum_{k=0}^{2^j-1} \hat{\gamma}_{0jk} W_{jk}(t/T), \quad t = 1, 2, \ldots, T. \tag{15}$$

7. For a given climate model $l \in \{1, \ldots, L\}$, generate B pairs of pseudo-sequences, $\{(\boldsymbol{X}^*_{bl}, \boldsymbol{Z}^*_b) : b = 1, \ldots, B\}$. The b-th pair contains a length-T pseudo-sequence derived from $\boldsymbol{X}_l$, denoted by $\boldsymbol{X}^*_{bl}$, and a length-T pseudo-sequence derived from $\boldsymbol{Z}$, denoted by $\boldsymbol{Z}^*_b$. To do this, create the bootstrapped values

$$\boldsymbol{X}^*_{bl} = (X^*_{bl}(1), \ldots, X^*_{bl}(T))', \text{ where } \quad X^*_{bl}(t) = \hat{\alpha}_l + \hat{\beta}_l t + \hat{\mu}_0(t) + R^*_{bl}(t), \tag{16}$$

$$\boldsymbol{Z}^*_b = (Z^*_b(1), \ldots, Z^*_b(T))', \text{ where } \quad Z^*_b(t) = \hat{\alpha}_0 + \hat{\beta}_0 t + \hat{\mu}_0(t) + R^*_{b0}(t), \tag{17}$$

where $R^*_{bl}(t)$ is the b-th simulated residual, $l = 0, 1, \ldots, L$. For the given l under consideration, note that the *same* values $\hat{\mu}_l(t) = \hat{\mu}_0(t)$ are used in Eqs. (16) and (17) thus enforcing the null hypothesis. To simulate $R^*_{bl}(t)$, see step 8.

8. Simulation of $R^*_{bl}(t)$, $t = 1, 2, \ldots, T$, $b = 1, 2, \ldots, B$:

   (a) Define $\boldsymbol{R}_l = (R_l(1), R_l(2), \ldots, R_l(T))'$ as the residual time series,

$$\boldsymbol{R}_l = \boldsymbol{X}_l - \hat{\alpha}_l \mathbf{1} - \hat{\beta}_l \mathbf{t} - \hat{\boldsymbol{\mu}}_0, \tag{18}$$

   where $\mathbf{1}$ is the column vector of one's of length T, and $\mathbf{t} = (1, 2, \ldots, T)'$. Fit an auto-regressive integrated moving average (ARIMA) model (Brockwell and Davis, 1991) to $\boldsymbol{R}_l$, and denote the fitted model by

$$A_l(\hat{\phi}_{l1}, \ldots, \hat{\phi}_{l(p+d)}, \hat{\theta}_{l1}, \ldots, \hat{\theta}_{l(lq)}, \hat{\tau}^2_l), \tag{19}$$

   where p and q are the numbers of parameters in the autoregressive and moving average components of the model, and d is the degree of differencing applied to make the time series $\boldsymbol{R}_l$ stationary. The estimated coefficients of the autoregressive part of the model are $\hat{\phi}_{l1}, \ldots, \hat{\phi}_{l(p+d)}$, and the estimated coefficients of the moving average part are $\hat{\theta}_{l1}, \ldots, \hat{\theta}_{lq}$. The estimate of the noise variance is τ^2_l.

   (b) Simulate the b-th realization from the fitted model $A_l(\hat{\phi}_{l1}, \ldots, \hat{\phi}_{l(p+d)}, \hat{\theta}_{l1}, \ldots, \hat{\theta}_{l(lq)}, \hat{\sigma}^2_l)$ by setting $R^*_{bl}(1) = R_{bl}(1)$, $R^*_{bl}(2) = R_{bl}(2)$,..., $R^*_{bl}(d) = R_{bl}(d)$, sampling $\epsilon^*_l(t)$ from $N(0, \tau^2_l)$, and computing

$$R^*_{bl}(t) = \hat{\phi}_{l1} R^*_{bl}(t-1) + \ldots + \hat{\phi}_{l(p+t-1)} R^*_{l(p+t-1)}(1-p) + \\ \hat{\theta}_{l1}(t-1)\epsilon_l(t-1) + \ldots + \hat{\theta}_{lq}\epsilon_l(t-q) + \epsilon^*_{bl}(t), \tag{20}$$

   for $t = d+1, d+2, \ldots, T$.

9. For $b = 1, \ldots, B$, and a given l, obtain the values D^*_{bl} from $\boldsymbol{X}^*_{bl}$ and $\boldsymbol{Z}^*_b$ as follows.

   (a) Obtain $\tilde{\tilde{\boldsymbol{X}}}^*_{bl}$ and $\tilde{\tilde{\boldsymbol{Z}}}^*_b$ by repeating Step 2 above with $\boldsymbol{X}^*_{bl}$ in place of $\boldsymbol{X}_l$ and $\boldsymbol{Z}^*_b$ in place of $\boldsymbol{Z}$.

   (b) Obtain $\tilde{\boldsymbol{X}}^*_{bl}$ and $\tilde{\boldsymbol{Z}}^*_b$ by repeating Step 3 above with $\tilde{\tilde{\boldsymbol{X}}}^*_{bl}$ in place of $\tilde{\tilde{\boldsymbol{X}}}_l$ and $\tilde{\tilde{\boldsymbol{Z}}}^*_b$ in place of $\tilde{\tilde{\boldsymbol{Z}}}$.

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

   (d) Compute the simulated values, $D^*_{bl} = \left(\hat{\boldsymbol{\Gamma}}^*_{bl} - \hat{\boldsymbol{\Gamma}}^*_{b0}\right)' \boldsymbol{\Omega} \left(\hat{\boldsymbol{\Gamma}}^*_{bl} - \hat{\boldsymbol{\Gamma}}^*_{b0}\right)$, for $b = 1, \ldots, B$.

For a given $l \in \{1, \ldots, L\}$, the set $\{D^*_{bl} : b = 1, 2, \ldots, B\}$ gives an empirical approximation to the null distribution of D_l under H_0.

3.3 Computing p-values

Recall from Eq. (14) that the value of the test statistic, computed using the actual time sequences $\boldsymbol{X}_l$ and $\boldsymbol{Z}$, is denoted by D_l, for a given l. The collection $\{D^*_{bl} : b = 1, 2, \ldots, B\}$ approximates the sampling distribution of D_l under the null hypothesis that $\boldsymbol{X}_l$ and $\boldsymbol{Z}$ share the same climate-signal, estimated from $\boldsymbol{Z}$ by $\hat{\boldsymbol{\mu}}_0$ in Eq. (15).

The quantile at D_l in the distribution of $\{D^*_{bl} : b = 1, \ldots, B\}$ is an empirical approximation to one minus the p-value of the test of the null hypothesis, $H_{0l} : \boldsymbol{\Gamma}_{\boldsymbol{X}_l} = \boldsymbol{\Gamma}_{\boldsymbol{Z}}$, under the conditions and assumptions described in Section 3. It is interpreted here as a probability-scale measure of compatibility between the test statistic's value and how extreme it is under the null hypothesis (Wasserstein and Lazar, 2016). To emphasize this interpretation, we refer to these p-values as "compatibilities" and denote them by c_l. Specifically, the compatibility associated with the test is estimated by

$$c_l = P^*(D^*_l > D_l | H_0) \equiv \frac{\#(D^*_{bl} > D_l)}{B}, \tag{21}$$

where P^* denotes a probability with respect to the empirical distribution $\{D^*_{bl} : b = 1, \ldots, B\}$.

4 Case study: Evaluating CMIP5 models using observations

In this section, we demonstrate our methodology described in the previous sections 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).

4.1 Data sources

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

4.1.1 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 to be assessed in these comparisons. In this example, we define statistical agreement between two time sequences as agreement between their climate-scale wavelet coefficients, where our definition of climate-scale is the three coarsest wavelet coefficient levels. This corresponds roughly to ten-year periodicity. We emphasize that this choice is made here only to illustrate our methodology, and others may wish to define climate-scale with a different choice of threshold separating climate-scale from weather-scale in the wavelet decomposition hierarchy.

We obtained a total of 139 time sequences of global monthly mean near-surface air temperature anomalies, generated by 44 CMIP5 models, from the KNMI Climate Data Explorer website (`https://climexp.knmi.nl/`). Climate Data Explorer allows on-the-fly aggregation, averaging, and renormalization of data sets with a simple menu-driven interface. We selected all time sequences available 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 1961 – 1990, as in Flato et al. (2013).

The collection of sequences produced by a given model is called an ensemble; some models produced just one ensemble member, while other produced as many as ten. 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 use in this case study is May 1918 through August 2003; a sequence of exactly 1024 months. Table 1 lists the 44 models used in this study, the modeling centers that are responsible for them, and the size of the models’ ensembles.

4.1.2 HadCRUT4 observations

Following Flato et al. (2013), we used the HadCRUT4 data set (Morice et al., 2012) as our 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 Morice et al. (2012). As with the model simulations, we used the KNMI Climate Data Explorer to

Table 1. 44 CMIP5 models used in this study.

Model	*Center*	*Members*	*Model*	*Center*	*Members*
ACCESS1-0	CSIRO-BOM (Australia)	1	GFDL-ESM2M	GFDL (USA)	1
ACCESS1-3	CSIRO-BOM (Australia)	3	GISS-E2-H p1	NASA GISS (USA)	1
BCC-CSM-1	Beijing Climate Center (PRC)	3	GISS-E2-H-CC p1	NASA GISS (USA)	6
BCC-CSM-1-M	Beijing Climate Center (PRC)	3	GISS-E2-R p1	NASA GISS (USA)	1
BNU-ESM	Beijing Normal Univ. (PRC)	1	GISS-E2-R-CC p1	NASA GISS (USA)	6
CanSM2	CCCMA (Canada)	5	HadGEM2-AO	NIMR/KMA (UK/Korea)	1
CCSM4	NCAR (USA)	6	HadGEM2-CC	MOHC/INPE (UK/Brazil)	1
CESM1-BGC	NCAR/DOE/NSF (USA)	1	HadGEM2-ES	MOHC/INPE (UK/Brazil)	4
CESM1-CAM5	NCAR/DOE/NSF (USA)	3	INMCM4	INM (Russia)	1
CESM1-CAM5-1-FV2	NCAR/DOE/NSF (USA)	4	IPSL-CM5A-LR	IPSL (France)	6
CESM1-FASTCHEM	NCAR/DOE/NSF (USA)	3	IPSL-CM5A-MR	IPSL (France)	3
CESM1-WACCM	NCAR/DOE/NSF (USA)	1	IPSL-CM5B-LR	IPSL (France)	1
CMCC-CESM	CMCC (Italy)	1	MIROC-ESM	MIROC (Japan)	3
CMCC-CM	CMCC (Italy)	1	MIROC-ESM-CHEM	MIROC (Japan)	1
CMCC-CMS	CMCC (Italy)	1	MIROC5	MIROC (Japan)	5
CNRM-CM5	CNRM (France)	10	MPI-ESM-LR	MPI (Germany)	3
CSIRO-Mk3-6-0	CSIRO (Australia)	10	MPI-ESM-MR	MPI (Germany)	3
EC-EARTH	EC-EARTH Consortium (Europe)	9	MPI-ESM-P	MPI (Germany)	2
FGOALS-g2	LASG (PRC)	5	MRI-CGM3	MRI (Japan)	3
FIO-ESM	FIO (PRC)	3	MRI-ESM1	MRI (Japan)	1
GFDL-CM3	GFDL (USA)	5	NorESM1-M	NCC (Norway)	3
GFDL-ESM2G	GFDL (USA)	3	NorESM1-ME	NCC (Norway)	1

obtain the monthly global average near-surface temperature anomalies for the period May 1918 thought August 2003, where the anomalies are computed relative to the average of the period 1961-1990. 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`.

4.2 Exploratory comparison

Figure 1 shows time sequence plots of the first ensemble member from each of 44 CMIP5 model, with a slightly smoothed version of HadCRUT4 observations (for better readability) superimposed. All sequences are truncated to the period May 1918 through August 2003, which provides a sequence of 1024 months that includes the periods covered by all models and by 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 model 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.

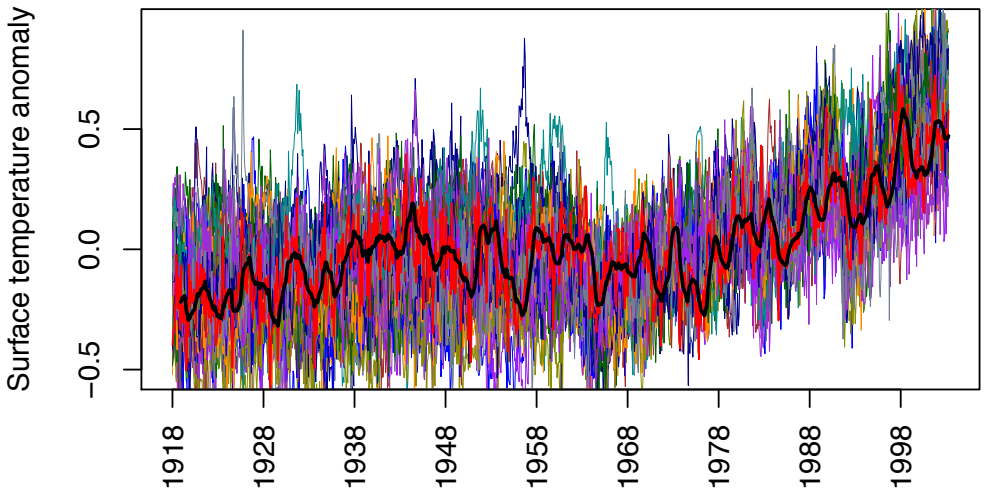

Figure 1. Monthly global average near-surface temperature anomaly time sequence plots for the first ensemble member of each of the 44 CMIP5 models (colors), and the HadCRUT4 observational sequence (red), May 1918–August 2003. The black line is a 12-month running mean computed from the HadCRUT4 data.

4.3 Compatibility of CMIP5 model simulations and observed HadCRUT4 data

We performed the steps described in Sections 3.2 and 3.3 on the 139 CMIP5 historical time sequences, using the HadCRUT4 observations as a benchmark. The number of replications in the simulation was set to $B = 5000$. No padding of the sequences was required since all time sequences are of length 1024. Padding may introduce artifacts by giving some time points in the sequences more importance than others, so it is desirable to avoid it if possible. The CMIP5 and HadCRUT4 data extend back to about 1850, but the HadCRUT4 observations are almost certainly less reliable as one goes back in time. For these reasons, we choose a period starting in the early 20th century and continuing for 1024 consecutive months, that covers the time period covered by all the models.

The DWT was applied to the detrended time sequences shown in Figure 2, with $\check{J} = 3$. This corresponds to a cycle of approximately ten years, which is, in our opinion, the finest time scale that one could legitimately call “climate". The choice of $\check{J}$ is important because it defines the set of temporal scales over which we evaluate agreement between models and observations. This choice 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. The choice of wavelet family did not affect our results significantly.

We used R's `wavethresh` (Nason, 2015) package for the wavelet decomposition, and the `forecast` package (Hyndman and Khandakar, 2008) for fitting the residual time sequences in step 8 of the procedure described in Section 3.2. In particular, forecast provides the `auto.arima` function, which automatically chooses the best ARIMA model by the AIC criterion. The base package's `arima.sim` function can then generate realizations from model fit by `auto.arima`, given its estimate of the noise variance, τ_l^2, and assuming that the residuals from the fit are a white noise process. To check the latter assumption, we ran the function `whitenoise.test` (Lobato and Velasco, 2004) from the package `normwhn.test`. Of the 139 time sequences, 23 failed this test: the null hypothesis of white noise was rejected at the 0.001 level. For these, we attempted to fit ARIMA models, possibly including seasonal components, manually. After re-checking the residuals, a total of nine ensemble members did not pass the white noise test on their residuals: CNRM-CM5/9, CSIRO-Mk3-6-0/7, CSIRO-Mk3-6-0/9, EC-EARTH/1, FGOALS-g2/1, FIO-ESM/2, GISS-E2-H-p1/1, GISS-E2-H-p1/3, and GISS-E2-R-p1/1. We proceeded with the processing of these sequences anyway, but note them as special cases in Figure 2 below.

Figure 2 displays the compatibilities, computed using the methodology of Section 3, for all 139 time sequences generated by the 44 models in the CMIP5 historical experiment. The models are arranged in alphabetical order along the x-axis of the graphic, and each ensemble member's compatibility value with the HadCRUT4 observations is shown by the vertical position of a plotting symbol. The nine time sequences for which the residuals from the ARIMA fit did not pass the white noise test, are indicated by asterisks. The figure shows a striking degree of variability among members produced by the same model. For example, the compatibilities of the ten time sequences generated by the CSIRO-Mk3-6-0 model with HadCRUT4, range from 0.0088 for member 10, to 0.9998 for member 5. To elucidate the correspondence between our results and model performance we now investigate CSIRO-Mk3-6-0 model's ensemble in greater depth.

Excluding CSIRO-Mk3-6-0/7 and CSIRO-Mk3-6-0/9 (due to failure of their residual sequences to pass the white noise test), the eight remaining members of the CSIRO-Mk3-6-0 ensemble are shown in Figure 3. The time sequences rendered in color are the best and worst performing members of the ensemble: members 5 and 10 respectively. The other six members are rendered in gray to give a general impression of their variability. Figure 4 shows the corresponding climate-signal time sequences after detrending, estimating the wavelet coefficients for the three coarsest levels of the wavelet decomposition, and transforming back to the time domain. For reference, the HadCRUT4 climate-signal, defined and computed in the same way, is superimposed in red. It's quite clear from this figure that the climate-signal time sequence for member 5 is closer to that of HadCRUT4 than is the climate-signal sequence for member 10. This is a reflection of the fact that the vector of climate-scale wavelet coefficients for member 5 is closer, in the metric D_l defined in Eq. (14), to the HadCRUT4 vector ($D_l = 0.305$ for member 5 versus 0.743 for member 10). This is only part of the story, however.

The other part of the story comes from the characteristics of the climate-noise time sequences that are left behind after accounting for trend and climate-signal. To obtain the null distribution of D_l that we require in order to understand the relative magnitudes of this quantity for the two members, we used a parametric bootstrap to create $B = 5000$ pairs of pseudo-realizations from a given ensemble member's time sequence and the observational time sequence. The bootstrapped observa-

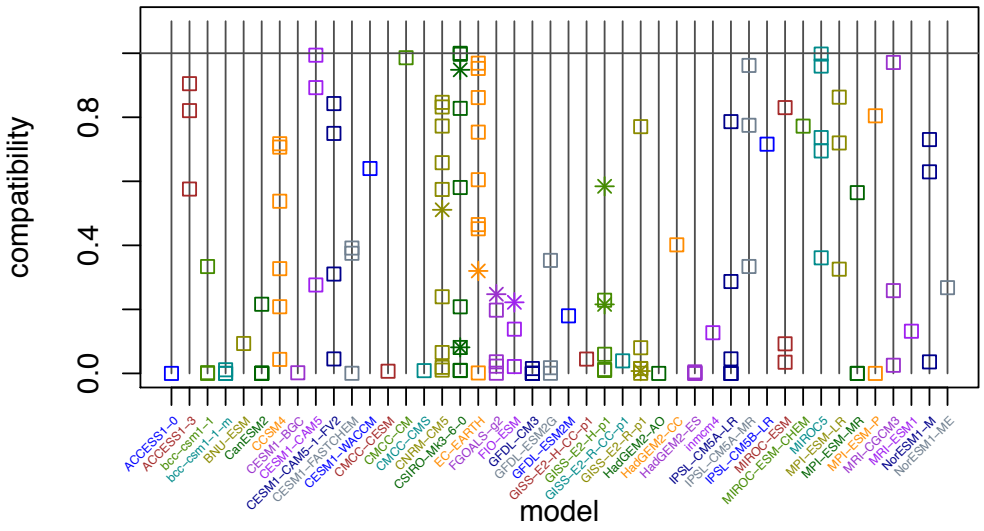

Figure 2. Model evaluation results for 139 time sequences generated by CMIP5 models in the historical experiment. Different models correspond to positions along the x-axis, with multiple ensemble members from the same model shown along the vertical line above the model name. Height along that line is the compatibility value. The maximum compatibility value of one is indicated by the gray horizontal line. Square plotting symbols indicate model ensemble members for which the residual from the ARIMA fit passed the white noise test. Asterisk plotting symbols indicated ensemble members which did not.

tional sequence is the sum of a) the HadCRUT4 trend, b) its climate-signal time sequence, and c) a bootstrapped realization from the following ARIMA$(1,0,1)$ which was fit to the HadCRUT4 climate-noise ($\boldsymbol{R}_0$ in Eq. (18)),

$$R_0(t) = \hat{\phi}_{01} R_0(t-1) + \hat{\theta}_{01}\epsilon_0(t-1) + \epsilon_0(t), \quad \epsilon_0(t) \sim N(0, \tau_0^2), \tag{22}$$

with

$$\hat{\phi}_{01} = 0.8539 \quad (\text{se}(\hat{\phi}_{01}) = 0.0243), \quad \hat{\theta}_{01} = -0.4223 \quad (\text{se}(\hat{\theta}_{01}) = 0.0422), \quad \hat{\tau}_0^2 = 0.0117. \tag{23}$$

The bootstrapped model sequence is the sum of a) the model's trend, b) the HadCRUT4 climate-signal time sequence, and c) a bootstrapped realization from the a time series model fit to the climate model's climate-noise ($\boldsymbol{R}_l$ in Eq. (18)). For CSIRO-Mk3-6-0/5 the best ARIMA model is ARIMA(1,0,2) with zero mean and coefficients,

$$\hat{\phi}_{l1} = 0.9390 \ (\text{se}(\hat{\phi}_{01}) = 0.0136), \ \hat{\theta}_{l1} = -0.3857 \ (\text{se}(\hat{\theta}_{l1}) = 0.0342), \ \hat{\theta}_{l2} = -0.0667 \ (\text{se}(\hat{\theta}_{l2}) = 0.0313), \ \hat{\tau}_0^2 = 0.009. \tag{24}$$

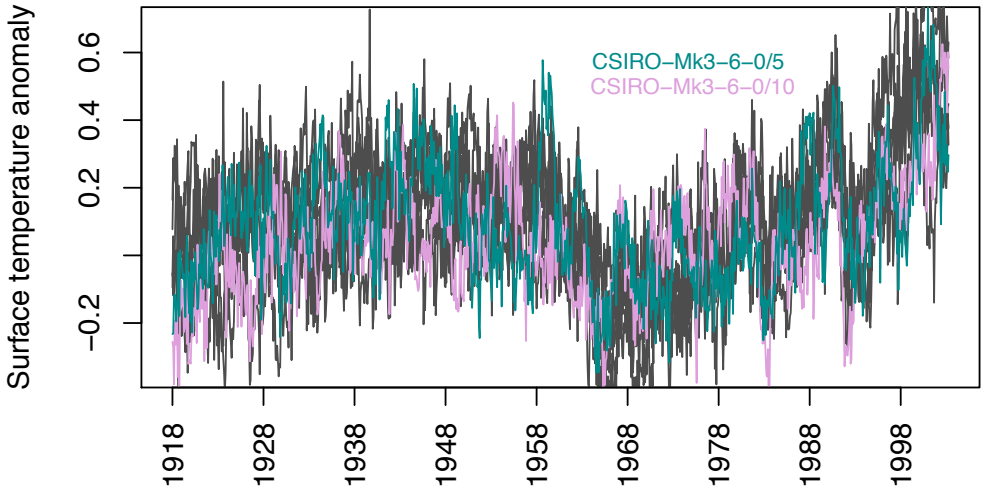

Figure 3. Time sequences of the CSIRO-Mk3-6-0 ensemble. The best and worst performing runs are members 5 and 10, respectively. They are shown in color. Members 1, 2, 3, 4, 6, and 8 are shown in grey. Members 7 and 9 are excluded from the analysis because they failed to meet required assumptions for ARIMA simulation.

For CSIRO-Mk3-6-0/10 the best ARIMA model is ARIMA(1,1,1) with coefficients,

$$\hat{\phi}_{l1} = 0.2107 \quad (\text{se}(\hat{\phi}_{01}) = 0.0643), \quad \hat{\theta}_{l1} = -0.6217 \quad (\text{se}(\hat{\theta}_{l1}) = 0.0514), \quad \hat{\tau}_0^2 = 0.008. \tag{25}$$

Thus, we have created 5000 bootstrapped realizations that mimic the statistical properties of HadCRUT4, and 5000 associated realizations that have the simulated climate signals of their HadCRUT4 companions, but trends and climate-noise sequences from the CSIRO-Mk3-6-0 member being evaluated (recall Eqs. (16) and (17)). Finally, each of 5000 companion pairs is evaluated as if they were new model runs paired with newly acquired observational data, yielding 5000 weighted squared distances between vectors of climate-scale wavelet coefficients.

Figure 5 illustrates this procedure. The two panels on the right show the original HadCRUT4 climate-signal time sequence in red (the same in both panels) and ten reconstructed climate-signal time sequences (in grey) out of the total of 5000, in each panel. Notice that HadCRUT simulated realizations used in the assessment of CSIRO-Mk3-6-0/5 (top-right) are different than the HadCRUT4 simulated realizations used in the assessment of CSIRO-Mk3-6-0/10 (bottom-right) because they are generated in separate simulations, even though they come from the same ARIMA model. The left panels show the same climate-signals from CSIRO-Mk3-6-0/5 (top-left) and CSIRO-Mk3-6-0/10 (top-right) that are displayed in Figure 4, plotted along with ten

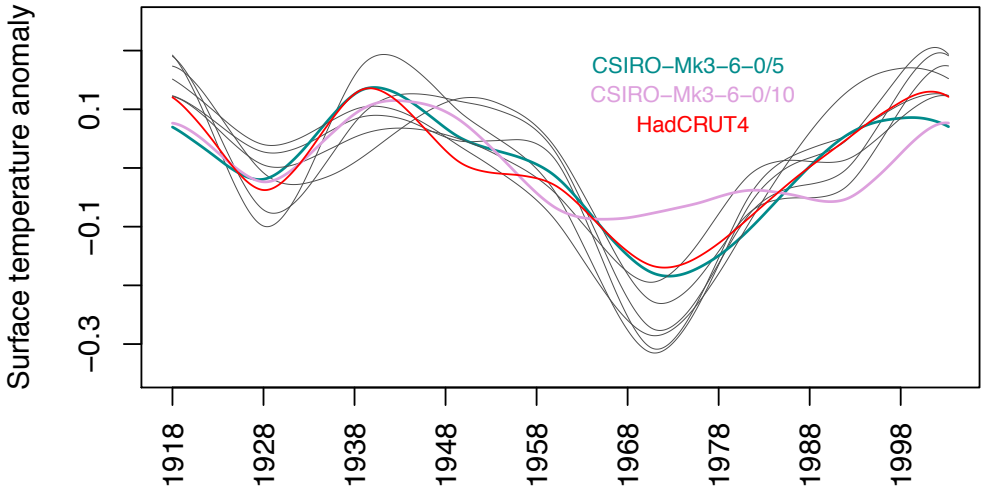

Figure 4. CSIRO-Mk3-6-0 ensemble members' (excluding members 7 and 9) climate-signal time sequences after detrending, estimating the wavelet coefficients for the three coarsest levels of the wavelet decomposition, and transforming back to the time domain. The HadCRUT4 climate-signal, defined and computed in the same way, is superimposed in red.

climate-signal time reconstructions (in grey) from the procedure described above. Every grey trajectory line in the left panels has a companion grey trajectory line in corresponding right panel. The similarities between these pairs of climate-scale time trajectories are quantified by the weighted squared distances between their climate-scale wavelet coefficients.

The second reason why CSIRO-Mk3-6-0/5 performs better in our evaluation than CSIRO-Mk3-6-0/10 is now evident: there is more variation in the climate-signal time sequences of member 5's bootstrapped realizations, than in member 10's. This is a consequence of differences in the structures of their climate-noise sequences; these structures are quantified by Eqs. (24) and Eqs. (25). Figure 6 is similar to Figure 5 except that it displays the climate-noise time sequences of the bootstrapped realizations instead of the climate-signal sequences. Greater variability in the noise portion of CSIRO-Mk3-6-0/5 relative to CSIRO-Mk3-6-0/10 must be a consequence of the difference in the two ARIMA models and their coefficients, and leads to more heterogeneity in its bootstrapped time sequences. This, in turn, leads to greater variability in the climate-signal wavelet coefficients of the bootstrapped time sequences derived from CSIRO-Mk3-6-0/5.

This conclusion is driven home in Figure 7. The right panel shows kernel density estimates, fit using R's density function, of the null distributions of the test statistic, D_l, for the eight members of the CSIRO-Mk3-6-0 ensemble under study. The left panel is identical except that only members 5 and 10 are colored (to make them easy to identify), and the actual values of

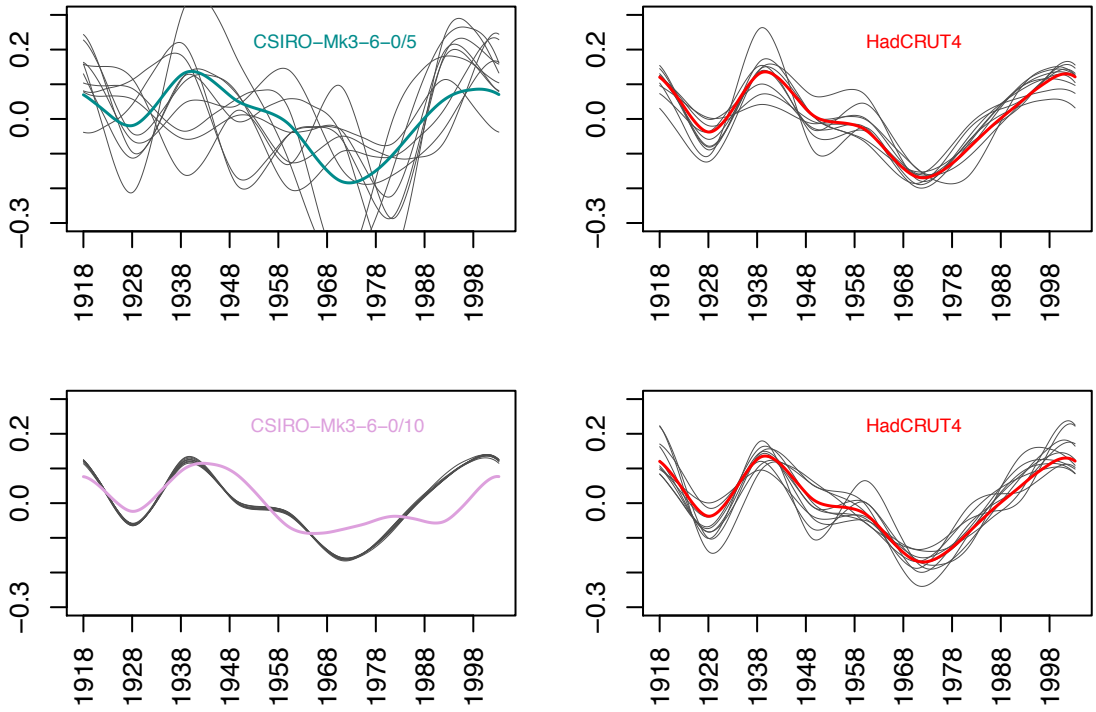

Figure 5. Impact of internal variability on bootstrapped climate signals. Top-right and bottom-right panels show the original HadCRUT4 climate-signal time sequence in red (the same in both panels), and ten reconstructed climate-signal time sequences (in grey). Top-left and bottom-left panels show the climate-signals from CSIRO-Mk3-6-0/5 and CSIRO-Mk3-6-0/10, respectively, along with reconstructed climate-signal sequences from ten bootstrapped realizations (in grey).

their respective test statistics are shown by suitably colored vertical lines. This makes clear that the dominant reason why the compatibility value for CSIRO-Mk3-6-0/5 is so high is the variability of its climate-noise time sequence. The right panel shows that the different ensemble members exhibit a variety of levels of this kind of internal variability.

For a single time sequence, generated either by a climate model or an observational data source, we regard climate-noise as a proxy for internal variability, and our method uses a parametric bootstrap to create pseudo-realizations from it. When added to the appropriate trend and climate-signal sequences, we thus create pseudo-realizations of full time sequences having the same statistical characteristics as their original counterparts. When uncertainties on observational data are not available, this may be a viable strategy for mimicking the aggregated effects of natural variability and observational error. When only a single member of a climate model ensemble exists, as is the case for some of the CMIP5 models in the historical experiment, the method may present a way of representing internal model variability. In fact, even when multiple ensemble members do

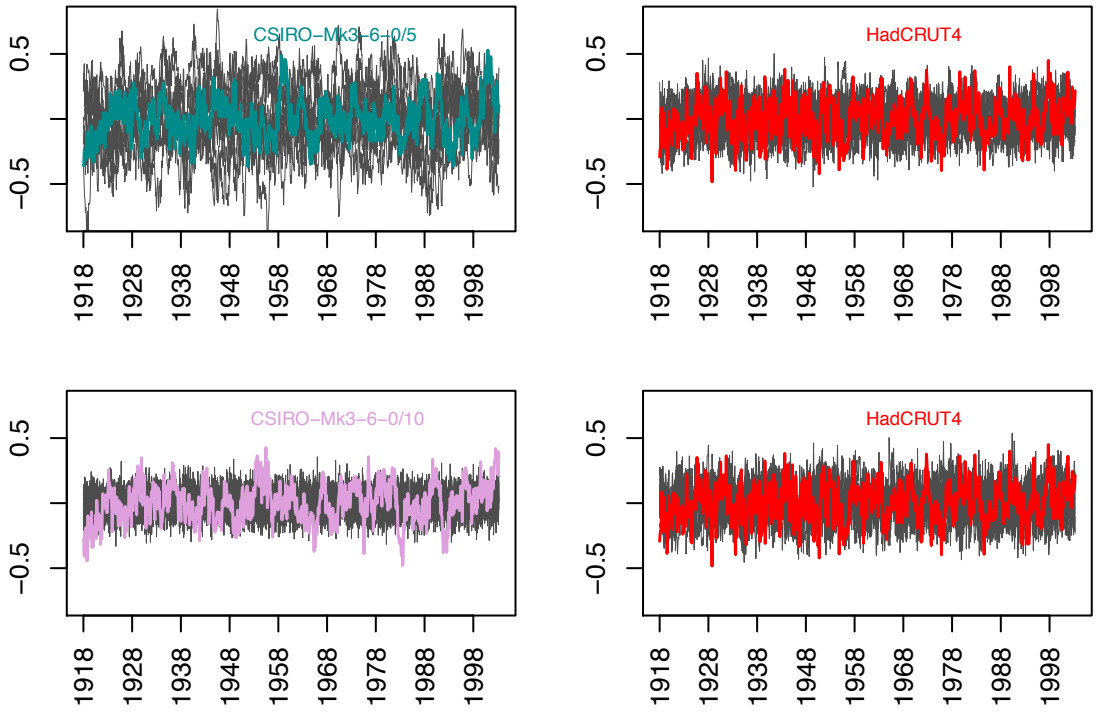

Figure 6. Climate-noise portions (in grey) of ten bootstrapped time sequences corresponding to the climate signals shown in Figure 5, with climate noise of the actual sequences superimposed.

exist, we argue that they are the results of purposeful perturbations of initial conditions and model parameters, and should be regarded as a source of “between" member variability rather than “within" member variability.

5 Conclusion

We have introduced a method, based on a hypothesis testing framework, to determine the degree to which climate-scale temporal-dependence structures in an observational time sequence are reproduced by climate-model-simulated time sequences. For a given climate model, the degree of agreement, or compatibility, is quantified by an empirical p-value from a test of the null hypothesis that climate-scale temporal dependence is the same in both the observed and climate-model-simulated time sequences. A p-value is the probability that a discrepancy as large or larger than that computed from the climate-model-simulated and observed sequences would be obtained, if the null hypothesis were true; that is, if the two sequences really did share the same climate-scale structure. In this context, a small empirical p-value suggests that a climate-signal in the climate model time sequence is incompatible with the climate-signal embedded in the observed time sequence.

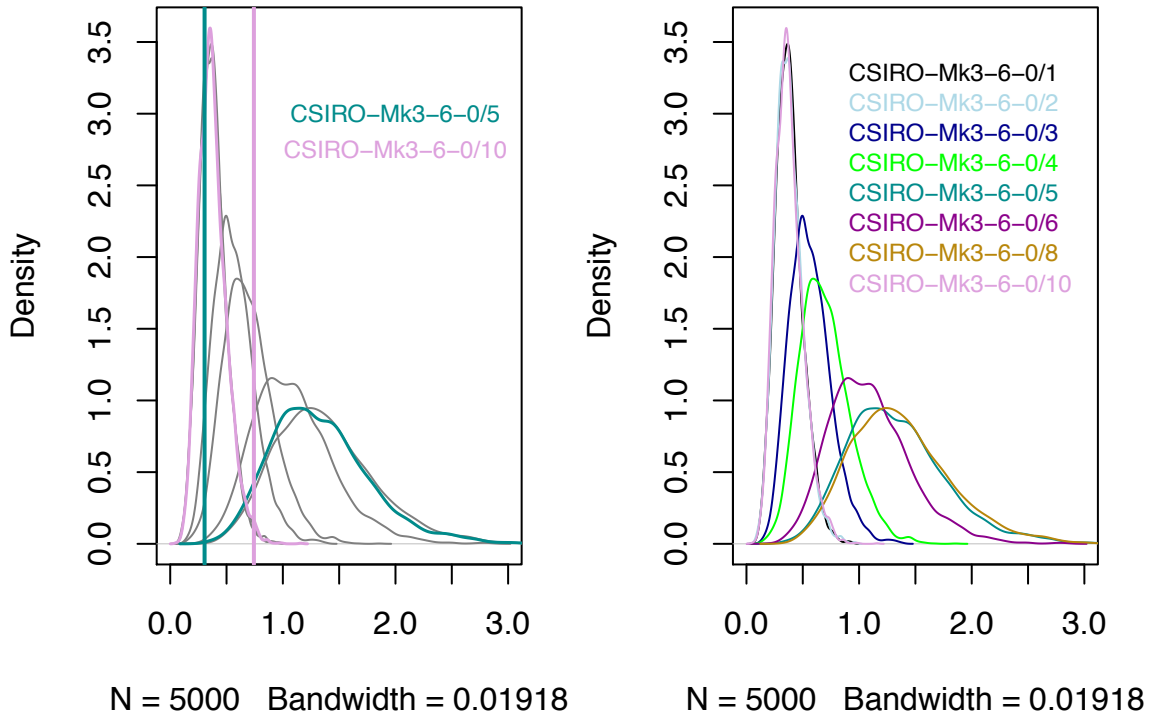

Figure 7. Null distributions, obtained by parametric bootstrapping, of eight members of the CSIRO-Mk3-6-0 model. Left panel: CSIRO-Mk3-6-0/5 and CSIRO-Mk3-6-0/10 highlighted, with their values of D_l indicated by the vertical lines. Right panel: Same as the left, but with the ensemble members identified by different colors.

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 necessary choices in this work that we believe to be reasonable, but others are certainly possible. The choice of the wavelet-decomposition level that constitutes the boundary between climate-signal and climate-noise is particularly important, since experiments have shown that it can change the results substantially. Users of this 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 our D_l are also possible and likely useful.

A crucially important methodological question about this approach is whether our strategy creates variabilities that are reasonable proxies for internal variabilities of a climate model, and of the natural climate system. It begs the question of what, exactly, “internal variability” means. We offer here an alternative, or perhaps a compliment, to the usual and somewhat

problematic definition that internal variability or uncertainty is captured by the spread of a multi-model or perturbed physics ensemble. At the very least, we hope this work will stimulate discussion on the topic.

Finally, there are natural extensions of this method to spatial and spatio-temporal contexts. Moving from one-dimensional to two-dimensional wavelets would allow us to use the same ideas on spatial 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 wavelet models may not be suitable in all cases. We are investigating the use of other basis functions and bootstrapping methods for these more complex settings.

6 Code availability

The code used in Section 4 will be made available either through JPL's open source mechanism or at the University of Minnesota.

7 Data availability

The data used in Section 4 are available through the KNMI Climate Data Explorer and is processed using code that will be made available either through JPL's open source mechanism or at the University of Minnesota.

Author contributions. The hypothesis testing strategy and formulation of compatibilities was a collaborative effort among all authors, as was the formulation of the statistical model that underlies the method. M. Heyman and S. Chatterjee developed the wavelet-based model for time sequences, and the bootstrapping framework for generating null distributions. A. Braverman carried out the analysis reported in Section 4. N. Cressie suggested the use of compatibilities and their justification as empirical p-values. A. Braverman prepared the manuscript with contributions from all co-authors.

Competing interests. None.

Disclaimer. None.

Acknowledgements. 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 and Advanced Information Systems Technology programs. In addition, Cressie's research was partially supported by a 2015–2017 Australian Research Council Discovery Grant, #DP150104576, and Chatterjee's research was partially supported by the National Science

Foundation (NSF) under grants # IIS-1029711 and # DMS-1622483. The authors would like to thank Huikyo Lee and Stephen Leroy for their thoughtful and thorough comments on this work.

References

Annan, J. and Hargeaves, J.: Reliability of the CMIP3 ensemble, Geophysical Research Letters, 37, doi:0.1029/2009GL041994, 2010.

Baumberger, C., Knutti, R., and Hadorn, G.: Building confidence in climate model projections: an analysis of inferences from fit, in: WIREs Climate Change, edited by Zorita, E. and Hulme, M., Wiley, doi:10.1002/wcc.454, 2017.

Boe, J. and Terray, L.: Can metric-based approaches really improve multi-model climate projections? The case of summer temperature change in France, Climate Dynamics, 45, 1913–1928, doi:10.1007/s00382-014-2445-5, 2015.

Brockwell, P. J. and Davis, R. A.: Time Series: Theory and Methods, Springer, 1991.

Covey, C., AchutaRao, K., Cubasch, U., Jones, P., Lambert, S., Mann, M., Phillips, T., and Taylor, K.: An overview of the results of the Coupled Model Intercomparison Project (CMIP), Global and Planetary Change, 37, 103–133, 2003.

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

Eyring, V., Gleckler, P., Heinze, C., Stouffer, R., Taylor, K., Balaji, V., Guilyardi, E., Joussaume, S., Kindermann, S., Lawrence, B., Meehl, G., Righi, M., and Williams, D.: Towards improved and more routine Earth system model evaluation in CMIP, Earth System Dynamics, Discussions, doi:10.5194/esd-2016-26, 2016.

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S., Cox, W. C. P., Driouech, F., Emori, S., Eyring, V., Forest, C., Gleckler, P., Guilyardi, E., Jakob, C., Kattsov, V., Reason, C., and Rummukainen, M.: Evaluation of Climate Models, in: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, edited by Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P., Cambridge University Press, 2013.

Giorgi, F. and Mearns, L.: Calculation of average, uncertainty range and reliability of regional climate changes from AOGCM simulations via the 'reliability ensemble averaging' (REA) method, Journal of Climate, 15, 1141–1158, 2002.

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

Hung, M.-P., Lin, J.-L., Wang, W., Kim, D., Shinoda, T., and Weaver, S.: MJO and convectively coupled equitorial waves simulated by CMIP5 climate models, Journal of Climate, 26, 6185–6214, doi:10.1175/JCLI-D-12-00541.1, 2013.

Hyndman, R. and Khandakar, Y.: Automatic Time Series Forecasting: The forecast Package for R, Journal of Statistical Software, 47, 3169–3181, doi:10.18637/jss.v027.i03, 2008.

Kiehl, J.: Overview of climate modeling, in: Frontiers of climate modeling, edited by Kiehl, J. and Ramanathan, V., Cambridge University Press, 2006.

Lin, Y. and Franzke, C.: 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, 2015.

Lobato, I. and Velasco, C.: A simple and general test for white noise, Econometric Society 2004 Latin American Meetings 112, Econometric Society, 2004.

Meehl, G., Goddard, L., Murphy, J., Stouffer, R., Boer, G., Danabasoglu, G., Dixon, K., Giorgetta, M., Greene, A., Hawkins, E., Hegerl, G., Karoly, D., Keenlyside, N., Kimoto, M., Kirtman, B., Navarra, A., Pulwarty, R., Smith, D., Stammer, D., and Stockdale, T.: Decadal prediction: can it be skillful?, Bulletin of the American Meteorological Society, doi:10.1175/BAMS2887.1, 2009.

Morice, C., Kennedy, J., N.A. Rayner, N., and Jones, P.: 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, 2012.

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

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

Ogden, T. R.: Essential wavelet for statistical applications and data analysis, Birkhauser, 1997.

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

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

Sanderson, B. and Knutti, R.: On the interpretation of constrained climate model ensembles, Geophysical Research Letters, 39, doi:10.1029/2012GL052665, 2012.

Suh, M.-S. and Oh, S.-G.: Development of new ensemble methods based on the performance skills of regional climate models over South Korea, Journal of climate, 25, 7067–7082, doi:10.1175/jcli-d-11-00457.1, 2012.

Taylor, K., Stouffer, R., and Meehl, G.: An Overview of CMIP5 and the Experiment Design, Bulletin of the American Meteorological Society, pp. 485–498, doi:10.1175/BAMS-D-11-00094.1, 2012.

Tebaldi, C. and Knutti, R.: 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, 2007.

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

Watanabe, M., Suzuki, T., O’ishi, R., Komuro, Y., Watanabe, S., Emori, S., Takemura, T., Chikira, M., Ogura, T., Sekiguchi, M., Takata, K., Yamzaki, D., Yokohata, T., Nozawa, T., Hasumi, H., Tatebe, H., and Kimoto, M.: Improved climate simulation by MIROC5: mean states, variabilty, and climate sensitivity, Journal of Climate, doi:10.1175/2010JCLI3679.1, 2010.