Comparing Climate Time Series. Part II: A Multivariate Test

Timothy DelSole\textsuperscript{1} and Michael K. Tippett\textsuperscript{2}

\textsuperscript{1}Department of Atmospheric, Oceanic, and Earth Sciences, George Mason University
\textsuperscript{2}Department of Applied Physics and Applied Mathematics, Columbia University, New York, New York

Correspondence: Timothy DelSole (tdelsole@gmu.edu)

Abstract. This paper proposes an objective criterion for deciding if two multivariate time series come from the same stochastic process. Importantly, the test accounts for correlations in both space and time. The basic idea is to fit each multivariate time series to a Vector Autoregressive (VAR) Model, and then test the hypothesis that the parameters of the two models are equal. In the special case of a first-order VAR model, the model is a Linear Inverse Model (LIM), and the test constitutes a difference-in-LIM test. One application of this test is to decide whether simulations from dynamical models are consistent observations. This test is applied to the question of whether climate models generate realistic variability of annual mean North Atlantic Sea Surface Temperature (NASST). Given the disputed origin of multidecadal variability in NASST (e.g., it could be forced by anthropogenic aerosols or arise naturally from internal variability), the time series are filtered in two different ways appropriate to the two driving mechanisms. In either case, only a few climate models out of three dozen are found to be consistent with the observed NASST. In fact, by comparing time series from every possible pair of climate models, it is shown that climate models not only differ from observations, but also differ significantly from each other, unless the models come from the same institution. These model-to-model and model-to-observation discrepancies raise considerable challenges for reaching a consensus on the relative role of internal variability and forcing in decadal fluctuations of NASST.

Copyright statement. TEXT
1 Introduction

A basic question in climate modeling is whether a given model realistically simulates observations. This question can be addressed only in a statistical sense, since climate variables contain unpredictable weather variability. In the special case of a single random variable and independent samples, this question can be addressed by applying standard tests of equality of distributions, such as the t-test, F-test, Kolmogorov-Smirnov test, etc. However, in many climate studies, multiple variables are of concern, and the associated time series are serially correlated.

The above question is particularly acute in the context of North Atlantic Sea Surface Temperature (NASST) variability. The North Atlantic is an area of enhanced decadal predictability and thus a prime candidate for skillful predictions on multi-year time scales (Kushnir, 1994; Griffies and Bryan, 1997; Marshall et al., 2001; Latif et al., 2004, 2006; Keenlyside et al., 2008). However, the dominant mechanisms of North Atlantic variability remain unclear. Some argue that North Atlantic variability is a manifestation of internal variability and that its multi-decadal swings may be misattributed to external forcing, leading to overestimates of climate sensitivity (DelSole et al., 2011; Tung and Zhou, 2013). Others argue that these multi-decadal swings are forced mostly by anthropogenic aerosols, with internal variability playing a weak role (Booth et al., 2012). However, simulations in which external forcing drives multi-decadal North Atlantic variability exhibit significant inconsistencies with observations, in terms of upper-ocean content, surface salinity, and spatial SST patterns (Zhang et al., 2013). Such discrepancies cast doubt on the claim that aerosols are a prime driver of twentieth-century North Atlantic variability. On the other hand, the alternative theory that North Atlantic multi-decadal variability is dominated by internal variability also exhibits inconsistencies. Specifically, climate models give inconsistent estimates of the magnitude and spatial structure of predictability in the North Atlantic (Branstator et al., 2012). However, the latter estimates are based on univariate analyses and not accompanied by a significance statement. As a result, it has proven difficult to reject a model based on its predictability or multi-year behavior, and models with widely different levels of predictability continue to remain equally plausible.

One of the most well-developed techniques for comparing time series is optimal fingerprinting (Bindoff et al., 2013). However, optimal fingerprinting is concerned mostly with detecting and attributing forced variability. Whether the variability that remains after removing forced variability is consistent with the model’s internal variability typically is assessed through the residual consistency test of Allen and Tett (1999), which is a test of total variance. As a result, each individual component could have the wrong variance yet the total sum could match observations. For this reason, the residual consistency check is not a stringent test of consistency of internal variability.

To advance this topic, we propose an objective criterion for deciding if two multivariate time series come from the same stochastic process. To simplify the problem, we consider only second-order stationary processes, in which the mean and covariance function are invariant to translations in time. Although non-stationarity is important in climate, a rigorous framework based on stationarity provides a starting point for comparing non-stationary processes. Also, climate models exhibit large biases, so any test of a difference in means likely would confirm such differences and not be surprising. Accordingly, our focus here is to test differences in covariance functions. Such a test is equivalent to testing equality of power spectra, since power spectra and covariance functions are each other’s Fourier Transforms.
One approach to testing equality of covariance functions is to express the function as a covariance matrix, and then apply standard tests of equality of covariance matrices (Anderson, 1984). However, these tests assume independent data and hence do not account for serial correlation. This paper proposes an approach to testing equality of covariance functions that accounts for serial correlation. Our basic idea is to fit data to a Vector Autoregressive Model (VAR) model, then test equality of parameters. The resulting test is the multivariate generalization of the test proposed by DelSole and Tippett (2020), which is part one of this paper series. Techniques for diagnosing differences in VAR models will be discussed in part three of this paper series.

2 Derivation of the Test

We first derive an exact test for equality of parameters for multivariate regression models, and then exploit asymptotic theory to derive a test for equality of multivariate AR processes.

We begin by considering the two multivariate regression models

\[ \begin{align*}
Y_1^* &= X_1^* B_1 + j_1 \mu_1^T + E_1 \\
Y_2^* &= X_2^* B_2 + j_2 \mu_2^T + E_2,
\end{align*} \]

where \( N_1 \) and \( N_2 \) are the respective sample sizes, \( S \) is the number of time series (variables), \( M \) the number of random predictors, and \( j_1 \) and \( j_2 \) are \( N_1 \)- and \( N_2 \)-dimensional vectors of ones. The dimensions of the above quantities are

\[ \begin{align*}
Y_1^* &\in \mathbb{R}^{N_1 \times S}, \quad X_1^* \in \mathbb{R}^{N_1 \times M}, \quad B_1 \in \mathbb{R}^{M \times S}, \quad j_1 \in \mathbb{R}^{N_1 \times 1}, \quad \mu_1 \in \mathbb{R}^{S \times 1} \\
Y_2^* &\in \mathbb{R}^{N_2 \times S}, \quad X_2^* \in \mathbb{R}^{N_2 \times M}, \quad B_2 \in \mathbb{R}^{M \times S}, \quad j_2 \in \mathbb{R}^{N_2 \times 1}, \quad \mu_2 \in \mathbb{R}^{S \times 1}.
\end{align*} \]

The matrices \( E_1 \) and \( E_2 \) are independent and distributed as

- rows of \( E_1 \) \( \overset{iid}{\sim} \mathcal{N}_S(0, \Gamma_1) \)
- rows of \( E_2 \) \( \overset{iid}{\sim} \mathcal{N}_S(0, \Gamma_2) \).

In this paper, we are interested in comparing variability, hence we test hypotheses without restricting the intercept terms \( \mu_1 \) and \( \mu_2 \). Accordingly, our null hypothesis is

\[ H_0 : \quad B_1 = B_2 \quad \text{and} \quad \Gamma_1 = \Gamma_2. \]

Let \( B_0 \) and \( \Gamma_0 \) denote the common regression parameters and noise covariance matrix under \( H_0 \), respectively.

It turns out that inferences based on models (1)-(2) are identical to inferences based on models

\[ \begin{align*}
Y_1 &= X_1 B_1 + E_1 \\
Y_2 &= X_2 B_2 + E_2,
\end{align*} \]

where the intercept terms \( \mu_1 \) and \( \mu_2 \) have been dropped and \( X, Y \) contain centered variables (i.e., the mean of each column is subtracted from that column). For simplicity then, we hereafter consider models (3)-(4), with the understanding that \( X, Y \)
refer to centered variables, and one extra degree of freedom is subtracted from each model to account for dropping $\mu_1$ and $\mu_2$.

It proves convenient to define

$$\lfloor \lfloor A \rfloor \rfloor = A^T A.$$  

Maximum Likelihood Estimates (MLEs) of the regression parameters are (Anderson, 1984, chapter 6)

$$\hat{B}_1 = (X_1^T X_1)^{-1} X_1^T Y_1 \quad \text{and} \quad \hat{B}_2 = (X_2^T X_2)^{-1} X_2^T Y_2,$$

the MLEs of the noise covariance matrices are

$$\Gamma_1 = \lfloor \lfloor Y_1 - X_1 \hat{B}_1 \rfloor \rfloor / N_1$$

and the associated maximized likelihoods are

$$l_1 = \frac{1}{(2\pi)^{SN_1/2}} \frac{1}{|\Gamma_1|^{N_1/2}} e^{-SN_1/2}$$

$$l_2 = \frac{1}{(2\pi)^{SN_2/2}} \frac{1}{|\Gamma_2|^{N_2/2}} e^{-SN_2/2}.$$  

Under $H_0$, the MLEs of $\mathbf{B}_0$ and $\Gamma_0$ are

$$\hat{B}_0 = (X_1^T X_1 + X_2^T X_2)^{-1} (X_1^T Y_1 + X_2^T Y_2)$$

and the associated maximized likelihood is

$$l_0 = \frac{1}{(2\pi)^{S(N_1+N_2)/2}} \frac{1}{|\Gamma_0|^{(N_1+N_2)/2}} e^{-S(N_1+N_2)/2}.$$  

It follows that the likelihood ratio statistic for testing $H_0$ is

$$\frac{l_0}{l_1 l_2} = \left( \frac{|\Gamma_1|^{N_1} |\Gamma_2|^{N_2}}{|\Gamma_0|^{N_1+N_2}} \right)^{1/2}.$$  

Before proceeding further, we pause to consider the fact that MLEs of covariances are biased. To correct this bias, we replace sample sizes by the degrees of freedom

$$\nu_1 = N_1 - M - 1 \quad \text{and} \quad \nu_2 = N_2 - M - 1,$$

which leads to the bias-corrected likelihood ratio

$$\Lambda_{total} = \left( \frac{|\Gamma_1|^{\nu_1} |\Gamma_2|^{\nu_2}}{|\Gamma_0|^{\nu_1+\nu_2}} \right)^{1/2}.$$  


where

\[ \hat{\Gamma}_1 = \frac{\|Y_1 - X_1 \hat{B}_1\|}{\nu_1} \]  \hspace{1cm} (10)

\[ \hat{\Gamma}_2 = \frac{\|Y_2 - X_2 \hat{B}_2\|}{\nu_2} \]  \hspace{1cm} (11)

\[ \hat{\Gamma}_0 = \frac{\|Y_1 - X_1 \hat{B}_0\| + \|Y_2 - X_2 \hat{B}_0\|}{\nu_1 + \nu_2} \]  \hspace{1cm} (12)

Regression models (3) and (4) involve a total of \(2S + 2MS + S(S + 1)\) parameters. The model under \(H_0\) involves \(2S + MS + S(S + 1)/2\) parameters. If \(H_0\) is true, then standard asymptotic theory indicates that the deviance statistic \(D(1,2)\) has a chi-squared distribution with degrees of freedom equal to the difference in the number of parameters between models:

\[ D(1,2) = -2 \log \Lambda_{\text{total}} \sim \chi^2_{MS+S(S+1)/2}. \]  \hspace{1cm} (13)

This distribution holds for both \(-2 \log \Lambda_{\text{total}}\) and \(-2 \log (l_0/l_1/l_2)\) because the two are asymptotically equivalent. A more accurate sampling distribution for \(D(1,2)\) can be derived using Monte Carlo techniques, but for our data, the resulting significance thresholds differ from that of (13) by less than 4%. Since the difference is small and does not affect any of our conclusions, the asymptotic distribution (13) is satisfactory for our purposes, and the Monte Carlo technique is not discussed further.

A Vector Autoregressive Model (VAR) for a random vector \(z_t\) is of the form

\[ z_t = A_1 z_{t-1} + \cdots + A_p z_{t-p} + \epsilon_t, \]

where \(A_1, \ldots, A_p\) are constant \(S \times S\) matrices, and \(\epsilon_t\) is a Gaussian white noise process with covariance \(\Gamma\). For a realization of length \(N\), the above model can be written in the form (3) or (4), or equivalently \(Y^T = B^T X^T + E^T\), using the identifications

\[ Y^T = \begin{pmatrix} z_{p+1} & z_{p+2} & \cdots & z_N \end{pmatrix} \in \mathcal{R}^{S \times (N-P)} \]

\[ B^T = \begin{pmatrix} A_1 & A_2 & \cdots & A_P \end{pmatrix} \in \mathcal{R}^{S \times SP} \]

\[ X^T = \begin{pmatrix} z_p & z_{p+1} & \cdots & z_{N-1} \\ z_{p-1} & z_p & \cdots & z_{N-2} \\ \vdots & \vdots & \ddots & \vdots \\ z_1 & z_2 & \cdots & z_{N-p} \end{pmatrix} \in \mathcal{R}^{SP \times (N-P)}. \]

The MLEs of \(B\) and \(\Gamma\), and the associated maximized likelihood, have been investigated extensively (see Brockwell and Davis, 1991; Lütkepohl, 2005; Box et al., 2008). For finite samples of serially correlated processes, the exact sampling distribution of these quantities are prohibitively complicated, even for Gaussian distributions (Brockwell and Davis, 1991, ch 6). However, for large sample sizes, these distributions are consistent with those derived from linear regression theory derived above (Brockwell and Davis, 1991, theorem 8.1.2, sec. 8.9). Accordingly, we use (13) to test equality of VAR models for large sample size.
We now apply our test to compare North Atlantic sea surface temperature (NASST) variability between models and observations. In particular, we focus on comparing multi-year internal variability. The question arises as to how to extract internal variability from observations. There is considerable debate about the magnitude of forced variability in this region, particularly the contribution due to anthropogenic aerosols (Booth et al., 2012; Zhang et al., 2013). Accordingly, we consider two possibilities: that the forced response is well represented by (1) a second-order polynomial or (2) a ninth-order polynomial over 1854-2018. These two assumptions will be justified shortly.

If NASST were represented on a typical $1^\circ \times 1^\circ$ grid, then the number of grid cells would far exceed the available sample size. Accordingly, some form of dimension reduction is necessary. Given our focus on multi-year predictability, we consider only large-scale patterns. Accordingly, we project annual-mean NASST onto the leading eigenvectors of the Laplacian over the Atlantic between $0^\circ - 60^\circ$N. These eigenvectors form an orthogonal set of patterns that can be ordered by a measure of length scale from largest to smallest. The first six Laplacian eigenvectors are shown in fig. 1 (these were computed by the method of DelSole and Tippett, 2015). The first eigenvector is spatially uniform. Projecting data onto the first Laplacian eigenvector is equivalent to taking the area-weighted average in the basin. In the case of SST, the time series for the first Laplacian eigenvector is merely an AMV index (AMV stands for “Atlantic Multidecadal Variability”). The second and third eigenvectors are dipoles that measure the large-scale gradient across the basin. Subsequent eigenvectors capture smaller scale patterns.

For model data, we use pre-industrial control simulations of SST from phase 5 of the Coupled Model Intercomparison Project (CMIP5 Taylor et al., 2012). Control simulations use forcings that repeat year after year. As a result, interannual variability in control simulations come from internal dynamical mechanisms, not from external forcing.

For observational data, we use version 5 of the Extended Reconstructed SST dataset (ERSSTv5 Huang et al., 2017). We consider only the 165-year period 1854-2018. We first focus on time series for the first Laplacian eigenvector, which we call the AMV index. The corresponding least squares fit to second- and ninth-order polynomials in time are shown in fig. 2. The second-order polynomial captures the secular trend toward warmer temperatures but otherwise has weak multidecadal variability. In contrast, the ninth-order polynomial captures both the secular trend and multidecadal variability. There is no consensus as to whether this multidecadal variability is internal or forced. Therefore, to account for either possibility, we analyze two sets of residuals: one in which a 2nd-order polynomial in time is removed from each Laplacian time series, and one in which a ninth-order polynomial in time is removed from each Laplacian time series. The residuals are called anomalies, and the anomalies for the first seven Laplacian eigenvectors are shown in fig. 3. In the case of removing a 2nd-order polynomial, the anomaly time series for Laplacian-1 contains marked multi-decadal variability. Comparing such anomalies between observations and control simulations implicitly assumes that this variability is internal variability. In the case of removing a 9th-order polynomial, little to no multidecadal variability is evident in the time series (see right column of fig. 3). Comparing such anomalies between observation and control simulations implicitly assumes that the variability captured by the 9th-order polynomial is externally forced and therefore removed before comparing to models. Performing our test on the two sets of anomalies allows us to test...
the sensitivity of our conclusions to assumptions about the source of multi-decadal variability. Note that only Laplacians 1 and 2 are sensitive to this assumption; time series for Laplacians three or higher are nearly the same in the two cases.

To match the dimensions of the observational time series, time series of length 165 years are drawn from the control simulations. Then, the time series are split in half (82 and 83 years), and time series from the earlier half are compared to those in the later half. The idea is that if a time series comes from the same source, then the null hypothesis is true and the method should detect a difference at the expected rate. Thirty-six CMIP5 models have PI control simulations of this length.

**Model Selection**

The process of deciding which variables to include in a regression model is called *model selection*. Here, model selection consists of choosing the lag $p$ and the number of Laplacian eigenvectors $S$. While numerous selection criteria exist for the case in which the $Y$-variables have been pre-selected and are not open to selection, such criteria do not apply straightforwardly to
Figure 2. AMV index from ERSSTv5 (thin grey), and polynomial fits to a second-order (thick black) and ninth-order (red) polynomial.

VAR models because choosing $S$ is tantamount to choosing both X- and Y-variables simultaneously. Recently, a criterion for selecting both X- and Y-variables was proposed by DelSole and Tippett (2020), called Mutual Information Criterion (MIC). This criterion is consistent with small-sample corrected versions of Akaike’s Information Criterion (AIC) for multivariate model selection. In terms of the regression model (3), the criterion is

$$\text{MIC}(X;Y) = N_1 \log \left( \frac{|\Gamma_1|}{|\Sigma_1|} \right) + \mathcal{P},$$

where $\Sigma_1$ is the MLE of the covariance matrix of $Y_1$ and

$$\mathcal{P} = N_1(N_1 + 1) \left( \frac{M + S}{N_1 - M - S - 2} - \frac{S}{N_1 - S - 2} - \frac{M}{N_1 - M - 2} \right).$$

The procedure is to select the subset of X- and Y-variables that minimize MIC. The VAR model selected by this criterion for each CMIP5 model is shown in table 1. MIC selects slightly different $p$ and $S$ for each CMIP5 model. In contrast, our method requires the same $p$ and $S$ for each VAR model. Therefore, a single compromise model must be chosen. Since the vast majority of selected VAR models are first order, we choose a first-order VAR, which yields a Linear Inverse Model (LIM). This
Figure 3. Projection of SST anomalies from ERSSTv5 onto the first seven Laplacian eigenvectors of the North Atlantic domain. The left and right panels show, respectively, anomalies derived by removing a second- and ninth-order polynomial in time. The integers on the right end of each time series indicates the Laplacian eigenvector.

choice is attractive because it allows us to make connection with the extensive literature on LIMs. In effect, our method is a “difference-in-LIM” test. (Incidentally, MIC provides a criterion for selecting the variables to include in LIMs.) Inspection of residuals from the LIMs show no obvious serial correlations, suggesting that the residuals are approximately white and that the first-order VAR model is adequate. We choose $S = 7$ Laplacian eigenvectors, which is the maximum for any first-order VAR. This choice is more likely to capture important dependency structures than a lower-order VAR model. This choice might also lead to overfitting, but one should remember that the distribution of the deviance statistic does not depend on the actual values of the regression parameters. Consequently, if some regression parameters vanish, indicating that the associated predictors are
Table 1. Models Selected by MIC, based on 82-year time series from each CMIP5 control simulation.

| Model                  | lags | laplacians | Model                  | lags | laplacians |
|------------------------|------|------------|------------------------|------|------------|
| GISS-E2-H              | 1    | 3          | CESM1-BGC              | 1    | 4          |
| GISS-E2-H-CC           | 1    | 3          | CESM1-CAM5             | 1    | 4          |
| GISS-E2-R              | 1    | 2          | CESM1-FASTCHEM         | 1    | 4          |
| GISS-E2-R-CC           | 1    | 3          | CESM1-WACCM            | 1    | 4          |
| MPI-ESM-LR             | 1    | 5          | CanESM2                | 1    | 6          |
| MPI-ESM-MR             | 1    | 4          | ACCESS1-0              | 1    | 4          |
| MPI-ESM-P              | 1    | 4          | ACCESS1-3              | 1    | 4          |
| CNRM-CM5               | 1    | 5          | CSIRO-Mk3-6-0          | 2    | 2          |
| CNRM-CM5-2             | 1    | 4          | inmcm4                 | 1    | 1          |
| CMCC-CESM              | 1    | 5          | IPSL-CM5A-LR           | 2    | 4          |
| CMCC-CM                | 1    | 4          | IPSL-CM5A-MR           | 1    | 4          |
| CMCC-CMS               | 2    | 3          | IPSL-CM5B-LR           | 1    | 6          |
| GFDL-CM3               | 1    | 7          | MIROC-ESM              | 1    | 4          |
| GFDL-ESM2G             | 1    | 5          | MIROC-ESM-CHEM         | 1    | 4          |
| GFDL-ESM2M             | 1    | 3          | HadGEM2-CC             | 1    | 7          |
| NorESM1-M              | 1    | 4          | HadGEM2-ES             | 1    | 5          |
| NorESM1-ME             | 1    | 4          | MRI-CGCM3              | 1    | 4          |
| CCSM4                  | 2    | 2          | ersst                  | 1    | 5          |

redundant and can be dropped, the test remains exactly the same as for a model in which all regression parameters are non-zero. The main adverse consequence of including more variables than necessary is a loss of power (i.e., the test is less able to detect a difference when such a difference exists). In this study, we are more concerned with capturing important dependency structures than with loss of power, hence our choice to include the maximum number of Laplacian eigenvectors selected by MIC.

To be clear, our method can be applied to arbitrary VAR models. Our choice of VAR model is merely a compromise designed to capture significant large-scale spatial dependencies while also adequately modeling temporal correlations.

Results

The deviance between ERSST 1854-1936 and 82-year segments of pre-industrial control simulations are shown in fig. 4. Also shown is the deviance between ERSST 1854-1936 and ERSST 1937-2018 (first item on x-axis). The latter deviance falls below the 5% threshold and hence indicates no significant difference between two halves of ERSST, regardless of polynomial fit. This result is consistent with the hypothesis that ERSST is a stationary VAR process after removing either a second- or ninth-order polynomial. Only one CMIP5 model is consistent with ERSST when a 2nd-order polynomial removed, and only two CMIP5
Figure 4. Deviance between ERSSTv5 1854-1935 and 82-year segments from 36 CMIP5 pre-industrial control simulations. Also shown is the deviance between ERSSTv5 1854-1935 and ERSSTv5 1937-2018 (first item on x-axis). The black and red curves show, respectively, results after removing a second- and ninth-order polynomial in time over 1854-2018 before evaluating the deviance. The models have been ordered on the x-axis from smallest to largest deviance after removing a second-order polynomial in time.

models are consistent with ERSST when a 9th-order polynomial is removed. We conclude that the vast majority of CMIP5 models do not generate realistic simulations of ERSST.

To explore the sensitivity of the above results to the number of Laplacians, deviances based on 10 Laplacian eigenvectors are shown in fig. 5. In this case, every CMIP5 model differs from ERSST, regardless of which polynomial is removed. This illustrates a general pattern: as the number of Laplacian eigenvectors increases, differences between stochastic processes become easier to detect. Because adding Laplacian eigenvectors corresponds to resolving smaller-scales, this pattern means that dis-
crepancies between models, and between models and observations, become more detectable as smaller-scale spatial structures are taken into account.

It is instructive to change the reference time series used for comparison. For instance, instead of comparing to ERSST, we compare each time series to time series from the CanESM2 model. The result of comparing every time series from the first half to every time series in the second half is summarized in fig. 6. The plotted numerical value is the deviance divided by its 5% critical value. Light and dark grey shading indicate significant differences at the 5% and 1%, respectively. Note that the diagonal is unshaded, indicating that the test correctly concludes no difference in VAR model when time series come from the same CMIP5 model. Some CMIP5 models differ significantly from all other models (e.g., MRI, inmcm4), indicating that these models not only are inconsistent with observations, but also inconsistent with other CMIP5 models.
Interestingly, models from the same institution tend to be indistinguishable from each other (e.g., GISS, NCAR, MPI, CMCC). This result is consistent with previous studies indicating that models developed at the same institution show more similarities with each other than with models developed at different institutes (Knutti et al., 2013). The deviances for 10 Laplacian vectors is shown in fig. 7. With more Laplacians comes the ability to detect more differences, so that almost all models are found to differ significantly from each other, unless they come from the same institution.

An alternative approach to summarizing dissimilarities between models is through dendograms (Knutti et al., 2013; Izenman, 2013). An example is shown in fig. 8 and constructed according to the following iterative procedure. First, each model is assigned to its own cluster. Next, the pair of VAR models with the smallest deviance are merged together to form a new cluster. This clustering is indicated by a two-pronged “leaf” whose length equals the deviance. After VAR models are linked, an “agglomeration” rule is used to measure dissimilarity relative to a cluster. There is no unique choice for the agglomeration rule. We choose a standard one called complete-linkage clustering, in which the “distance” between two sets of clusters $A$ and $B$ is defined to be the maximum deviance between elements of each cluster:

$$\max \{\text{deviance}(x, y) : x \in A, y \in B\}. $$

If this distance exceeds the significance threshold, then at least one deviance between elements of the clusters is known to be significant. Then, the smallest distance between any pair of clusters is linked together again. This process repeats until all models have been merged into a single cluster. The x-axis shows the distance and the y-axis shows the data sources. Each source name occurs twice because two time series are drawn from that source. As can be seen, the shortest leaves are associated with time series from the same source, or from models developed at the same institution.

The broad conclusions drawn from the dendogram in fig. 8 are similar to those drawn by Knutti et al. (2013). However, the dendogram constructed in Knutti et al. (2013) was based on the Kullback-Leibler (KL) divergence. Importantly, KL divergence requires estimating covariance matrices, but a significance test for this measure was not examined in Knutti et al. (2013). The dendogram developed here is attractive in that it is based on a rigorous statistical measure of dissimilarity that has a well-defined significance threshold that accounts for sampling variability and dependencies in space and time.

A perennial question is whether models should be weighted equally when making multi-model projections of the future. Such weighting schemes lie outside the scope of this paper, but a related question is whether there exists a relation between a model’s past performance and its predictions of the future. To investigate this question, we plot a model’s deviance from ERSST against that model’s Equilibrium Climate Sensitivity (ECS). ECS is the equilibrium change in annual mean global surface temperature following a doubling of atmospheric CO2 concentration. The result is shown in fig. 9. The figure also shows a least-square line fit to the data points. The slope is statistically significant ($p = 0.011$) under the standard assumptions of independent and identically distributed residuals, which of course are dubious for our problem. Nevertheless, the correlation is negative, indicating that models that best simulate the past tend to have larger ECS. Extrapolating to zero deviance yields an ECS of 4.8°C, although the probabilistic meaning (e.g., confidence interval) of this result is unclear.
This paper proposed an approach to deciding if two multivariate time series come from the same stochastic process. The basic idea is to fit each time series to a Vector Autoregressive Model, and then test if the parameters of the models are equal. The likelihood ratio test for this problem and the associated sampling distributions were derived. This derivation leads to a deviance statistic that measures the difference between VAR processes and can be used to rank models based on their “closeness” to the VAR process inferred from observations. The test accounts for correlations in time and correlations between variables. In the special case of a first-order VAR model, the model is a LIM and the test is effectively a “difference-in-LIM” test.

The test was illustrated by using it to compare annual mean North Atlantic SST variability in models and observations. When compared to observations, almost every CMIP5 model differs significantly from ERSST. This conclusion holds regardless of whether a second- or ninth-order polynomial in time is regressed out. Thus, our conclusion does not depend on whether multidecadal NASST variability is assumed to be forced or internal. By applying a hierarchical clustering technique, we showed that time series from the same model, or from models from the same institution, tend to be clustered together, and in many cases differ significantly from other clusters. Our results are consistent with previous claims (Pennell and Reichler, 2011; Knutti et al., 2013) that the effective number of independent models is smaller than the actual number of models in a multi-model ensemble.

Recently, Mann et al. (2021) argued that there exists no compelling evidence for internal multidecadal oscillations. In particular, oscillations seen in proxies of pre-industrial temperature can be explained as an artifact of volcanic activity that happens to project on the multidecadal frequency band. Under this hypothesis, variability in NASST is due primarily to external forcing, and removing a ninth-order polynomial would remove forced variability better than a second-order polynomial. Nevertheless, even after removing a ninth-order polynomial, our results show that most models are inconsistent with observations. Thus, work remains to be done to improve model simulations of internal variability.

Given the differences that are detected, it is of interest to diagnose the spatial-temporal structure of those differences. Methods for diagnosing such structure will be discussed in part three of this series of papers.
1. Series are obtained by extracting a continuous 165-year period, regressing out a second-order polynomial, then splitting the time series in half (22 and 33 years). For observations, the 165-year period corresponds to 1851-2018. The deviance is divided by the 5% significance thresholds, so values greater than one indicate a significant difference in VAR model. Light and dark grey shading highlight values greater than the 5% and 1% significance thresholds. White spaces indicate insignificant differences between VAR models.

Figure 6. The deviance between two non-overlapping time series from CMIP5 pre-industrial control simulations and observations. The time series are obtained by extracting a continuous 165-year period, regressing out a second-order polynomial, then splitting the time series in half (22 and 33 years). For observations, the 165-year period corresponds to 1851-2018. The deviance is divided by the 5% significance thresholds, so values greater than one indicate a significant difference in VAR model. Light and dark grey shading highlight values greater than the 5% and 1% significance thresholds. White spaces indicate insignificant differences between VAR models.

### Table: Deviance Matrix for Total

|                | GISS-H | GISS-H-CC | GISS-R | GISS-R-CC |
|----------------|--------|-----------|--------|-----------|
| HadGEM2-CC     | 0.4    | 0.6       | 0.8    | 0.3       |
| HadGEM2-ES     | 0.5    | 0.7       | 0.9    | 0.4       |
| CMCC-CESM      | 0.6    | 0.8       | 1.0    | 0.5       |
| CMCC-ESM       | 0.7    | 0.9       | 1.1    | 0.6       |
| GFDL-CM2       | 0.8    | 1.0       | 1.2    | 0.7       |
| GFDL-ESM2G     | 0.9    | 1.1       | 1.3    | 0.8       |
| CNRM-CM5-2     | 1.0    | 1.2       | 1.4    | 0.9       |
| CNRM-CM5-2     | 1.1    | 1.3       | 1.5    | 1.0       |
| MIROC-ESM      | 1.2    | 1.4       | 1.6    | 1.1       |
| MIROC-CHEM     | 1.3    | 1.5       | 1.7    | 1.2       |
| MPI-LR         | 1.4    | 1.6       | 1.8    | 1.3       |
| MPI-MR         | 1.5    | 1.7       | 1.9    | 1.4       |
| MPI-P          | 1.6    | 1.8       | 2.0    | 1.5       |
| NorESM1-M      | 1.7    | 1.9       | 2.1    | 1.6       |
| NorESM1-ME     | 1.8    | 2.0       | 2.2    | 1.7       |
| CCSM4          | 1.9    | 2.1       | 2.3    | 1.8       |
| CESM1-BGC      | 2.0    | 2.2       | 2.4    | 1.9       |
| CESM1-CAM5     | 2.1    | 2.3       | 2.5    | 2.0       |
| FASTCHEM       | 2.2    | 2.4       | 2.6    | 2.1       |
| WACCM          | 2.3    | 2.5       | 2.7    | 2.2       |
| CanESM2        | 2.4    | 2.6       | 2.8    | 2.3       |
| ACCESS1-0      | 2.5    | 2.7       | 2.9    | 2.4       |
| ACCESS1-3      | 2.6    | 2.8       | 3.0    | 2.5       |
| CSIRO          | 2.7    | 2.9       | 3.1    | 2.6       |
| INMCM4         | 2.8    | 3.0       | 3.2    | 2.7       |
| IPSL-A-LR      | 2.9    | 3.1       | 3.3    | 2.8       |
| IPSL-MR        | 3.0    | 3.2       | 3.4    | 2.9       |
| IPSL-B-LR      | 3.1    | 3.3       | 3.5    | 3.0       |
| MRI            | 3.2    | 3.4       | 3.6    | 3.1       |

### Figure: Deviance Matrix for Total

The deviance between two non-overlapping time series from CMIP5 pre-industrial control simulations and observations. The time series are obtained by extracting a continuous 165-year period, regressing out a second-order polynomial, then splitting the time series in half (22 and 33 years). For observations, the 165-year period corresponds to 1851-2018. The deviance is divided by the 5% significance thresholds, so values greater than one indicate a significant difference in VAR model. Light and dark grey shading highlight values greater than the 5% and 1% significance thresholds. White spaces indicate insignificant differences between VAR models.
Figure 7. Same as fig. 6, but for 10 Laplacian eigenvectors.
Figure 8. Dendogram derived from the deviance matrix between all possible VAR models estimated from the first and second halves of the 1854-2018 period (the specific year is not relevant for pre-industrial control simulations). The clusters are agglomerated according to the complete-linkage clustering, which uses the maximum deviance between elements of each cluster. The VAR models contain seven Laplacian eigenfunctions and a second-order polynomial in time is removed. The vertical red line shows the 5% significance threshold for a significance difference in VAR model.
Figure 9. Deviance versus Equilibrium Climate Sensitivity of CMIP5 models. The deviance is computed for NASST separately for the first and second halves of the 1854-2018 period, which yields two points per CMIP5 model for a total of 72 points. ECS is derived from table 9.5 of Flato et al. (2013). The red line shows the least-squares line fit, and the red triangle at the top shows the intercept of the best-fit line.
Code availability. Upon publication, an R code for performing the statistical test described in this paper will be available at http://cola.gmu.edu/delsole/webpage/Overview/DelSoleCodeOverview.html

Author contributions. Both authors participated in the writing and editing of the manuscript. TD performed the numerical calculations.

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

Acknowledgements. This research was supported primarily by the National Oceanic and Atmospheric Administration (NA20OAR4310401). The views expressed herein are those of the authors and do not necessarily reflect the views of these agencies.
References

Allen, M. R. and Tett, S. F. B.: Checking for model consistency in optimal fingerprinting, Clim. Dyn., 15, 419–434, 1999.
Anderson, T. W.: An Introduction to Multivariate Statistical Analysis, Wiley-Interscience, 1984.
Bindoff, N. L., Stott, P. A., AchutaRao, K. M., Allen, M. R., Gillett, N., Gutzler, D., Hansingo, K., Hegerl, G., Hu, Y., Jain, S., Mokhov, I. I., Overland, J., Perlwitz, J., Webbari, R., and Zhang, X.: Detection and Attribution of Climate Change: From Global to Regional, 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., Qin, D., Plattner, G.-K., Tignor, M., Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P., chap. 10, pp. 867–952, Cambridge University Press, 2013.
Booth, B. B., Dunstone, N. J., Halloran, P. R., Andrews, R., and Bellouin, N.: Aerosols implicated as a prime driver of twentieth-century North Atlantic climate variability, Nature, 484, 228–232, 2012.
Box, G. E. P., Jenkins, G. M., and Reinsel, G. C.: Time Series Analysis: Forecasting and Control, Wiley-Interscience, 4th edn., 2008.
Branstator, G., Teng, H., Meehl, G. A., Kimoto, M., Knight, J. R., Latif, M., and Rosati, A.: Systematic estimates of initial value decadal predictability for six AOGCMs, J. Climate, 25, 1827–1846, 2012.
Brockwell, P. J. and Davis, R. A.: Time Series: Theory and Methods, Springer Verlag, 2nd edn., 1991.
DelSole, T. and Tippett, M. K.: Laplacian Eigenfunctions for Climate Analysis, J. Climate, 28, 7420–7436, 2015.
DelSole, T. and Tippett, M. K.: Comparing climate time series – Part 1: Univariate test, Adv. Stat. Clim. Meteorol. Oceanogr., 6, 159–175, https://doi.org/10.5194/ascmo-6-159-2020, https://ascmo.copernicus.org/articles/6/159/2020/, 2020.
DelSole, T. and Tippett, M. K.: A Mutual Information Criterion with Applications to Canonical Correlation Analysis and Graphical Models, STAT, in press, 2020.
DelSole, T., Tippett, M. K., and Shukla, J.: A significant component of unforced multidecadal variability in the recent acceleration of global warming, J. Climate, 24, 909–926, 2011.
Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S., Collins, W., Cox, 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., Qin, D., Plattner, G.-K., Tignor, M., Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P., pp. 741–866, Cambridge University Press, 2013.
Griffies, S. M. and Bryan, K.: A predictability study of simulated North Atlantic multidecadal variability, Clim. Dyn., 13, 459–487, 1997.
Huang, B., Thorne, P. W., Banzon, V. F., Boyer, T., Chepurin, G., Lawrimore, J. H., Menne, M. J., Smith, T. M., Vose, R. S., and Zhang, H.-M.: Extended Reconstructed Sea Surface Temperature, Version 5 (ERSSTv5): Upgrades, Validations, and Intercomparisons, Journal of Climate, 30, 8179–8205, https://doi.org/10.1175/JCLI-D-16-0836.1, https://doi.org/10.1175/JCLI-D-16-0836.1, 2017.
Izenman, A. J.: Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning, Springer, corrected 2nd printing edn., 2013.
Keenlyside, N. S., Latif, M., Jungclaus, J., Kornblueh, L., and Roeckner, E.: Advancing decadal-scale climate prediction in the North Atlantic sector, Nature, 453, 84–88, 2008.
Knutti, R., Masson, D., and Gettelman, A.: Climate model genealogy: Generation CMIP5 and how we got there, Geophysical Research Letters, 40, 1194–1199, https://doi.org/10.1002/grl.50256, https://doi.org/10.1002/grl.50256, 2013.
Kushnir, Y.: Interdecadal variations in the North Atlantic sea surface temperature and associated atmospheric conditions, J. Climate, 7, 141–157, 1994.

Latif, M., Roeckner, E., Botzet, M., Esch, M., Haak, H., Hagemann, S., Jungclaus, J., Legutke, S., Marsland, S., Mikolajewicz, U., and Mitchell, J.: Reconstrucing, Monitoring, and Predicting Multidecadal-Scale Changes in the North Atlantic Thermohaline Circulation with Sea Surface Temperature, J. Climate, 17, 1605–1614, 2004.

Latif, M., Collins, M., Pohllmann, H., and Keenlyside, N.: A review of predictability studies of Atlantic sector climate on decadal time scales, J. Climate, 19, 5971–5987, 2006.

Lütkepohl, H.: New introduction to multiple time series analysis, Spring-Verlag, 2005.

Mann, M. E., Steinman, B. A., Brouillette, D. J., and Miller, S. K.: Multidecadal climate oscillations during the past millennium driven by volcanic forcing, Science, 371, 1014–1019, https://doi.org/10.1126/science.abc5810, https://science.sciencemag.org/content/371/6533/1014, 2021.

Marshall, J., Kushnir, Y., Battisti, D., Chang, P., Czaja, A., Dickson, R., Hurrell, J., McCartney, M., Saravanan, R., and Visbeck, M.: North Atlantic Climate Variability: Phenomena, Impacts, and Mechanisms, Int. J. Climatology, 21, 1863–1898, 2001.

Pennell, C. and Reichler, T.: On the Effective Number of Climate Models, Journal of Climate, 24, 2358 – 2367, https://doi.org/10.1175/2010JCLI3814.1, https://journals.ametsoc.org/view/journals/clim/24/9/2010jcli3814.1.xml, 2011.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An Overview of CMIP5 and the Experimental Design, Bull. Am. Meteor. Soc., 93, 485–498, 2012.

Tung, K.-K. and Zhou, J.: Using data to attribute episodes of warming and cooling in instrumental records, Proceedings of the National Academy of Sciences, 110, 2058–2063, 2013.

Zhang, R., Delworth, T. L., Sutton, R., Hodson, D. L. R., Dixon, K. W., Held, I. M., Kushnir, Y., Marshall, J., Ming, Y., Msadek, R., Robson, J., Rosati, A. J., Ting, M., and Vecchi, G. A.: Have Aerosols Caused the Observed Atlantic Multidecadal Variability?, J. Atmos. Sci., 70, 1135–1144, 2013.