Adv. Stat. Clim. Meteorol. Oceanogr., 6, 159–175, 2020
https://doi.org/10.5194/ascmo-6-159-2020

Comparing climate time series – Part 1: Univariate test

Timothy DelSole[1] **and Michael K. Tippett**[2]

[1]Department of Atmospheric, Oceanic, and Earth Sciences, George Mason University,
Fairfax, Virginia, USA
[2]Department of Applied Physics and Applied Mathematics, Columbia University, New York, New York, USA

Correspondence: Timothy DelSole (tdelsole@gmu.edu)

Received: 9 February 2020 – Revised: 18 August 2020 – Accepted: 25 August 2020 – Published: 12 October 2020

Abstract. This paper proposes a new approach to detecting and describing differences in stationary processes. The approach is equivalent to comparing auto-covariance functions or power spectra. The basic idea is to fit an autoregressive model to each time series and then test whether the model parameters are equal. The likelihood ratio test for this hypothesis has appeared in the statistics literature, but the resulting test depends on maximum likelihood estimates, which are biased, neglect differences in noise parameters, and utilize sampling distributions that are valid only for large sample sizes. This paper derives a likelihood ratio test that corrects for bias, detects differences in noise parameters, and can be applied to small samples. Furthermore, if a significant difference is detected, we propose new methods to diagnose and visualize those differences. Specifically, the test statistic can be used to define a "distance" between two autoregressive processes, which in turn can be used for clustering analysis in multi-model comparisons. A multidimensional scaling technique is used to visualize the similarities and differences between time series. We also propose diagnosing differences in stationary processes by identifying initial conditions that optimally separate predictable responses. The procedure is illustrated by comparing simulations of an Atlantic Meridional Overturning Circulation (AMOC) index from 10 climate models in Phase 5 of the Coupled Model Intercomparison Project (CMIP5). Significant differences between most AMOC time series are detected. The main exceptions are time series from CMIP models from the same institution. Differences in stationary processes are explained primarily by differences in the mean square error of 1-year predictions and by differences in the predictability (i.e., R-square) of the associated autoregressive models.

1 Introduction

Climate scientists often confront questions of the following types.

1. Does a climate model realistically simulate a climate index?
2. Do two climate models generate similar temporal variability?
3. Did a climate index change its variability?
4. Are two power spectra consistent with each other?

Each of the above questions requires deciding whether two time series come from the same stochastic process. Although numerous papers in the weather and climate literature address questions of the above types, the conclusions often are based on visual comparison of estimated auto-covariance functions or power spectra without a rigorous significance test. Lund et al. (2009) provide a lucid review of some objective tests for deciding whether two time series come from the same stationary process. An additional test that was not considered by Lund et al. (2009) is to fit autoregressive models to time series and then to test differences in parameters (Maharaj, 2000; Grant and Quinn, 2017). Grant and Quinn (2017) showed that this test has good power and performs well even when the underlying time series are not from an autoregressive process. The latter test has been applied to such problems as speech recognition but not to climate time series. The purpose of this paper is to further develop this test for climate applications.

The particular tests proposed by Maharaj (2000) and Grant and Quinn (2017) test equality of autocorrelation without regard to differences in variances. However, in climate applications, differences in variance often are of considerable importance. In addition, these tests employ a sampling distribution derived from asymptotic theory and therefore may be problematic for small sample sizes. In this paper, the likelihood ratio test is derived for the more restrictive case of equality of noise variances, which leads to considerable simplifications, including a test that is applicable for small sample sizes. Further comments about how our proposed test compares with previous tests will be discussed below.

If the time series are deemed to come from different processes, then it is desirable to characterize those differences in meaningful terms and to group time series into clusters. Piccolo (1990) and Maharaj (2000) propose such classification procedures. Following along these lines, our hypothesis test suggests a natural measure for measuring the distance between two stationary processes that can be used for clustering analysis. For multi-model studies, this distance measure can be used to give a graphical summary of the similarities and differences between time series generated by different models. In addition, we extend the interpretation of such summaries considerably by proposing a new approach to diagnosing differences in stationary processes based on finding the initial condition that optimally separates one-step predictions.

2 Previous methods for comparing time series

In this section we review previous methods for comparing two time series. These methods are based on the theory of stochastic processes and assume that the joint distribution of the values of the process at any collection of times is multivariate normal and stationary. Although non-stationarity is a prominent feature in many climate time series, a stationary framework is a natural starting point for non-stationary generalizations. Stationarity implies that the expectation at any time is constant, and the second-order moments depend only on the difference times (more precisely, this is called weak-sense stationarity). Accordingly, if X_t is a stationary process, then the mean is independent of time,

$$\mathbb{E}[X_t] = \mu_X \quad \text{(a constant)}, \tag{1}$$

and the time-lagged auto-covariance function depends only on the difference in times,

$$\mathbb{E}[(X_{t+\tau} - \mu_X)(X_t - \mu_X)] = c_X(\tau). \tag{2}$$

Because a multivariate normal distribution is fully characterized by its first and second moments, the stochastic process is completely specified by μ_X and the auto-covariance function $c_X(\tau)$. Stationarity further implies that the auto-covariance function is an even function of the lag τ. Following standard practice, we consider discrete time series where values are available at N_X equally spaced time steps $X_1, X_2, \ldots, X_{N_X}$.

Now consider another stationary process Y_t, with mean μ_Y and auto-covariance function $c_Y(\tau)$. The problem we consider is this: given sample time series $X_1, \ldots, X_{N_X}$ and $Y_1, \ldots, Y_{N_Y}$, decide whether the two time series come from the same stationary process. For stationary processes, we often are not concerned with differences in means (e.g., in climate studies, these often are eliminated through "bias corrections"), hence we allow $\mu_X \neq \mu_Y$. Therefore, our problem is equivalent to deciding whether the two time series have the same auto-covariance function, i.e., deciding whether

$$c_X(\tau) = c_Y(\tau) \quad \text{for all } \tau = 0, 1, \ldots. \tag{3}$$

This problem can be framed equivalently as deciding whether two stationary processes have the same *power spectrum*. Recall that the power spectrum is the Fourier transform of the auto-covariance function. We define the power spectrum of X_t as

$$p_X(\omega) = \sum_{\tau=-\infty}^{\infty} c_X(\tau) e^{i\omega\tau}. \tag{4}$$

The spectrum of Y_t is defined similarly and denoted as $p_Y(\omega)$. Because the Fourier transform is invertible, equality of auto-covariances is equivalent to equality of spectra. Thus, our problem can be framed equivalently as deciding whether

$$p_X(\omega) = p_Y(\omega) \quad \text{for all } \omega \in [0, \pi). \tag{5}$$

Estimates of the power spectrum are based on the *periodogram* (Box et al., 2008).

The above hypothesis differs from hypotheses about a single process that are commonly tested with auto-covariance functions or power spectra. For instance, the hypothesis of vanishing auto-correlation often is assessed by comparing the sample auto-correlation function to $\pm 1.96/\sqrt{N}$, where N is the length of the time series (e.g., Brockwell and Davis, 2002, chap. 1). In the spectral domain, the hypothesis of white noise often is tested based on the Kolmogorov–Smirnov test, in which the standardized cumulative periodogram is compared to an appropriate set of lines (e.g., Jenkins and Watts, 1968, Sect. 6.3.2). These tests consider hypotheses about *one* process. In contrast, our hypothesis involves a comparison of *two* processes.

Coates and Diggle (1986) derived *spectral-domain* tests for equality of stationary processes. The underlying idea of these tests is that equality of power spectra implies that their ratio is independent of frequency and therefore indistinguishable from white noise. This fact suggests that standard tests for white noise can be adapted to periodogram *ratios*. Coates and Diggle (1986) derive a second parametric test that assumes that the log ratio of power spectra is a low-order polynomial of frequency.

Lund et al. (2009) explored the above tests in the context of station data for temperature. They found that spectral methods have relatively low statistical power – that is, the methods are unlikely to detect a difference in stationary processes when such a difference exists. Our own analysis using the data discussed in Sect. 5 is consistent with Lund et al. (2009) (not shown). The fact that spectral-domain tests have less statistical power than time-domain tests is not surprising. After all, spectral tests are based on comparing periodograms in which the number of unknown parameters grows with sample size. In particular, for a time series of length N, the periodogram combines coefficients for sines and cosines of a Fourier transform into $N/2$ coefficients for the amplitude. Thus, a time series of length $N = 64$ yields a periodogram with 32 amplitudes; a time series of length $N = 512$ yields a periodogram with 256 amplitudes; and so on. Because the number of unknowns grows with sample size, the sampling error of the individual periodogram estimates does not decrease with sample size. Typically, sampling errors are reduced by *smoothing* periodogram estimates over frequencies, but the smoothing makes implicit assumptions about the shape of the underlying power spectrum. In the absence of hypotheses to constrain the power spectra, the large number of estimated parameters results in low statistical power.

Lund et al. (2009) also develop and discuss a *time-domain* test. This test is based on the sample estimate of the auto-covariance function

$$\hat{c}_X(\tau) = \frac{1}{N_X} \sum_{t=1}^{N_X - |\tau|} \left(X_{t+\tau} - \hat{\mu}_X\right)\left(X_t - \hat{\mu}_X\right), \qquad (6)$$

where $\hat{\mu}_X$ is the sample mean of X_t based on sample size N_X. The analogous estimate for the auto-covariance function of Y_t is denoted $\hat{c}_Y(\tau)$. The test is based on differences in auto-covariance functions up to lag τ_0. That is, the test is based on

$$\boldsymbol{\Delta} = \begin{pmatrix} \hat{c}_X(0) - \hat{c}_Y(0) \\ \hat{c}_X(1) - \hat{c}_Y(1) \\ \vdots \\ \hat{c}_X(\tau_0) - \hat{c}_Y(\tau_0) \end{pmatrix}. \qquad (7)$$

In the case of equal sample sizes $N_X = N_Y = N$, Lund et al. (2009) propose the statistic

$$\chi^2 = \frac{N}{2} \boldsymbol{\Delta}^T \widehat{\mathbf{W}}^{-1} \boldsymbol{\Delta}, \qquad (8)$$

where $\widehat{\mathbf{W}}$ is an estimate of the covariance matrix between auto-covariance estimates, namely

$$\hat{W}_{i+1,j+1} = \sum_{k=-K}^{K} \left(\hat{c}(k)\hat{c}(k-i+j) + \hat{c}(k+j)\hat{c}(k-i)\right), \qquad (9)$$

and $\hat{c}(\tau)$ is the pooled auto-covariance estimate

$$\hat{c}(\tau) = \frac{\hat{c}_X(\tau) + \hat{c}_Y(\tau)}{2}. \qquad (10)$$

The reasoning behind this statistic is lucidly discussed in Lund et al. (2009) and follows from standard results in time series analysis (see proposition 7.3.2 in Brockwell and Davis, 1991). Under the null hypothesis of equal auto-covariance functions, and for large N, the statistic χ^2 has an approximate chi-square distribution with $\tau_0 + 1$ degrees of freedom. A key parameter in this statistic is the cutoff parameter K in the sum for $\widehat{\mathbf{W}}$. Lund et al. (2009) propose using $K = N^{1/3}$ but acknowledge that this rule needs further study.

We have applied Lund et al.'s test to the numerical examples discussed in Sect. 5 and find that $\widehat{\mathbf{W}}$ is *not always positive definite*. In such cases, χ^2 is not guaranteed to be positive and therefore does not have a chi-square distribution. The lack of positive definiteness sometimes can be avoided by choosing a slightly different K, but in many of these cases the resulting χ^2 depends sensitively on K. This sensitivity arises from the fact that $\widehat{\mathbf{W}}$ is close to singular, so changing K by one unit can change χ^2 by more than an order of magnitude. Conceivably, some modification of $\widehat{\mathbf{W}}$ or some rule for choosing K can remove this sensitivity to K, but without such modification this test is considered unreliable. Further comments about this are given at the end of Sect. 5.

3 Comparing autoregressive models

Several authors have proposed approaches to comparing stationary processes based on assuming that the time series come from autoregressive models of order p, denoted AR(p). Accordingly, we consider the two AR(p) models

$$X_t = \phi_1^X X_{t-1} + \phi_2^X X_{t-2} + \ldots + \phi_p^X X_{t-p} + \gamma_X + \epsilon_t^X, \qquad (11)$$

$$Y_t = \phi_1^Y Y_{t-1} + \phi_2^Y Y_{t-2} + \ldots + \phi_p^Y Y_{t-p} + \gamma_Y + \epsilon_t^Y, \qquad (12)$$

where the ϕs are autoregressive parameters, the γs are constants that control the mean, and the ϵ_ts are independent Gaussian white noise processes with zero mean and variances

$$\sigma_X^2 = \mathrm{var}[\epsilon_t^X] \quad \text{and} \quad \sigma_Y^2 = \mathrm{var}[\epsilon_t^Y]. \qquad (13)$$

By construction, X_t and Y_s are independent for all t and s. Our method can be generalized to handle correlations between X_t and Y_s, specifically by using a vector autoregressive model with coupling between the two processes, but this generalization will not be considered here. For the above models, there exists a one-to-one relation between the first $p+1$ auto-covariances $c(0), \ldots, c(p)$ and the parameters $\phi_1, \ldots, \phi_p, \sigma^2$. This relation is expressed through the Yule–Walker equations and a balance equation for variance (Box et al., 2008). Therefore, equality of the first $p+1$ auto-covariances is equivalent to equality of the AR(p) parameters. Furthermore, estimates of the parameters depend only on the first $p+1$ auto-covariances. The remaining auto-covariances of an AR(p) process $c(p+1), c(p+2), \ldots$ are obtained through recursion formulas that depend only on the first $p+1$ auto-covariances. Importantly, the auto-covariance function of an AR(p) process does not depend on γ; rather,

Table 1. Definition of the variables in the test statistic $D_{\phi,\sigma}$ (Eq. 17).

$F_\sigma = \frac{\hat{\sigma}_X^2}{\hat{\sigma}_Y^2}$
$F_{\phi\vert\sigma} = \frac{(\widehat{\boldsymbol{\phi}}_X - \widehat{\boldsymbol{\phi}}_Y)^T \boldsymbol{\Sigma}_{\mathrm{HM}} (\widehat{\boldsymbol{\phi}}_X - \widehat{\boldsymbol{\phi}}_Y)}{p\hat{\sigma}^2}$
$\hat{\sigma}^2 = \frac{\nu_X \hat{\sigma}_X^2 + \nu_Y \hat{\sigma}_Y^2}{\nu_X + \nu_Y}$
$\widehat{\boldsymbol{\phi}}_X = \begin{pmatrix} \hat{\phi}_1^X \\ \hat{\phi}_2^X \\ \vdots \\ \hat{\phi}_p^X \end{pmatrix}$ and $\widehat{\boldsymbol{\phi}}_Y = \begin{pmatrix} \hat{\phi}_1^Y \\ \hat{\phi}_2^Y \\ \vdots \\ \hat{\phi}_p^Y \end{pmatrix}$
$\boldsymbol{\Sigma}_{\mathrm{HM}}$ defined in Eq. (A40)

γ controls the mean of the process. In climate studies, the mean often is removed from time series before any analysis, hence we do not require the means to be equal; i.e., we allow $\gamma_X \neq \gamma_Y$.

In previous tests, different authors make different assumptions about the noise variance. Maharaj (2000) allows the noise in the two models to differ in their variances and to be correlated at zero lag. Grant and Quinn (2017) allow the noise variances to differ but assume the noises are independent at zero lag. In the climate applications we have in mind, differences in variance are important, hence we are interested in detecting differences in noise variances. Accordingly, our null hypothesis of equivalent stationary processes is the following:

$$H_0: \left\{\phi_1^X = \phi_1^Y, \quad \ldots, \quad \phi_p^X = \phi_p^Y\right\} \quad \text{and } \sigma_X = \sigma_Y. \qquad (14)$$

The alternative hypothesis is that at least one of the above parameters differs between the two processes. Constraining processes to be AR(p) means that each process is characterized by $p+1$ parameters, in contrast to the unconstrained case in which the auto-covariance function or power spectrum is characterized by an infinite number of parameters.

The likelihood ratio test for hypothesis H_0 is derived in the Appendix. The derivation differs from that of Grant and Quinn (2017) primarily by including the hypothesis $\sigma_X = \sigma_Y$ in the null hypothesis, which leads to considerable simplifications in estimation and interpretation. Furthermore, we propose a modification to correct for biases and a Monte Carlo technique for determining significance thresholds. We describe the overall procedure here and direct the reader to the Appendix for details. The test relies on sample estimates of the above parameters, so we define these first. The autoregressive parameters in Eq. (11) are estimated using the method of least squares, yielding the least squares estimates $\hat{\phi}_1^X, \ldots, \hat{\phi}_p^X, \hat{\gamma}^X$. (The least squares method arises when maximizing the *conditional* likelihood, which is conditional on the specific realization of $X_1, \ldots, X_p$ in the data. For large sample sizes, the conditional least squares estimates will be close to the familiar maximum likelihood estimates.) The noise variance for X_t is estimated using the unbiased estimator

$$\hat{\sigma}_X^2 = \frac{\sum_{t=p+1}^{N_X} \left(X_t - \hat{\phi}_1^X X_{t-1} - \ldots - \hat{\phi}_p^X X_{t-p} - \hat{\gamma}^X\right)^2}{\nu_X}, \qquad (15)$$

where $\nu_X = N_X - 2p - 1$ is degrees of freedom, computed as the difference between the sample size $N_X - p$ and the number of estimated parameters $p+1$. This noise variance estimate is merely the residual variance of the AR model and is readily available from standard time series analysis software. Similarly, the method of least squares is used to estimate the AR parameters for Y_t in Eq. (12), yielding the least squares estimates $\hat{\phi}_1^Y, \ldots, \hat{\phi}_p^Y, \hat{\gamma}^Y$ and the noise variance estimate

$$\hat{\sigma}_Y^2 = \frac{\sum_{t=p+1}^{N_Y} \left(Y_t - \hat{\phi}_1^Y Y_{t-1} - \ldots - \hat{\phi}_p^Y Y_{t-p} - \hat{\gamma}^Y\right)^2}{\nu_Y}, \qquad (16)$$

where $\nu_Y = N_Y - 2p - 1$. Then, the test of H_0 is based on the statistic

$$D_{\phi,\sigma} = D_\sigma + D_{\phi\vert\sigma}, \qquad (17)$$

where

$$\begin{aligned} D_\sigma = {} & \nu_X \log(\nu_X + \nu_Y / F_\sigma) + \nu_Y \log(\nu_X F_\sigma + \nu_Y) \\ & - (\nu_X + \nu_Y) \log(\nu_X + \nu_Y) \end{aligned} \qquad (18)$$

$$D_{\phi\vert\sigma} = (\nu_X + \nu_Y) \log\left(1 + \frac{p F_{\phi\vert\sigma}}{\nu_X + \nu_Y}\right), \qquad (19)$$

and the remaining variables are defined in Table 1. Under the null hypothesis H_0, it is shown in the Appendix that F_σ and $F_{\phi\vert\sigma}$ have the following approximate distributions:

$$F_\sigma \sim F_{\nu_X, \nu_Y}, \qquad (20)$$

$$F_{\phi\vert\sigma} \sim F_{p, \nu_X + \nu_Y}. \qquad (21)$$

Furthermore, the two statistics are *independent*. Critical values for D_σ and $D_{\phi\vert\sigma}$ are obtained *individually* from the critical values of the F-distribution, taking care to use the correct one-tailed or two-tailed test (see Appendix, particularly Eq. A31). In principle, the exact sampling distribution of $D_{\phi,\sigma}$ can be derived analytically because $D_{\phi,\sigma}$ is the sum of two random variables with known distributions. However, this analytic calculation is cumbersome, whereas the quantiles of $D_\sigma + D_{\phi\vert\sigma}$ can be estimated accurately and quickly by Monte Carlo techniques. Essentially, one draws random samples from F_{ν_X,ν_Y} and $F_{p,\nu_X+\nu_Y}$, substitutes these into Eqs. (18)–(19), evaluates $D_{\phi,\sigma}$ in Eq. (17), and then repeats this many times (e.g., 10 000 times). Note that the required repetitions do not grow with sample size N_X and N_Y

since the F-distributions can be sampled directly. The 5 % significance threshold for $D_{\phi,\sigma}$, denoted D_{crit}, is then obtained from the 95th percentile of the Monte Carlo samples. Although a Monte Carlo technique is used to obtain critical values, the test still constitutes a test for small sample sizes.

The above test assumes that time series come from an AR(p) process *of known order* and are excited by Gaussian noise. Grant and Quinn (2017) use Monte Carlo simulations to show that the test is robust to non-Gaussianity. Since our proposed method assumes normal distributions through Eqs. (20) and (21), its robustness to departures to Gaussian remains an open question and is a topic for future study. If the order of the process is unknown and has to be selected or the time series does not come from an autoregressive model (e.g., the process is a moving average process), then the type I error rate does not match its nominal value. In such cases, Grant and Quinn (2017) propose using some sufficiently large value of p, such as

$$p^* = \lfloor (\log \min[N_X, N_Y])^\nu \rfloor, \tag{22}$$

where $\lfloor k \rfloor$ denotes the largest integer smaller than or equal to k. The resulting AR(p^*) model can capture the first p^*+1 auto-covariances regardless of their true origin. Of course, there will be some loss of power when a test based on p^* is applied to a time series that comes from an AR(p) model with $p < p^*$. This criterion seems to work well in our examples, as discussed in more detail in Sect. 5.

4 Interpretation of differences in AR processes

If a difference in the AR process is detected, then we would like to interpret those differences. After all, such comparisons often are motivated to validate climate models, so if a discrepancy is found we would like to describe those differences in meaningful terms. The fact that the statistic $D_{\phi,\sigma}$ can be expressed as the sum of two *independent* terms suggests that it is natural to consider the two terms separately. The first term D_σ depends only on the difference in noise variance estimates $\hat{\sigma}_X^2$ and $\hat{\sigma}_Y^2$. The noise variance σ_X^2 is the *one-step prediction error of the AR model*. After all, if the AR parameters in Eq. (11) were known exactly, then the conditional mean would be

$$\begin{aligned} S_X =& \mathbb{E}\left[X_t \mid X_{t-1}, X_{t-2}, \ldots, X_{t-p}\right] = \phi_1^X X_{t-1} \\ &+ \phi_2^X X_{t-2} + \ldots - \phi_p^X X_{t-p} + \gamma_X, \end{aligned} \tag{23}$$

where “S” stands for “signal”, and the one-step prediction error would be

$$X_t - \mathbb{E}\left[X_t \mid X_{t-1}, \ldots, X_{t-p}\right] = \epsilon_t^X. \tag{24}$$

Comparing the variance of one-step prediction errors is more statistically straightforward than comparing variances of the original time series because prediction errors are approximately uncorrelated, since they are the residuals of the AR model, which acts as a *pre-whitening transformation*. In contrast, comparing variances of time series is not straightforward because of serial correlation. In practice, the residuals are only approximately uncorrelated because the parameter values are only approximate.

The second term $D_{\phi|\sigma}$ vanishes when $\widehat{\boldsymbol{\phi}}_X = \widehat{\boldsymbol{\phi}}_Y$ and is positive otherwise, hence it clearly measures the difference in AR parameters. To further interpret this term, it is helpful to partition the AR model for X_t into two parts, an unpredictable part associated with the noise ϵ_t^X and a predictable part S_X. The predictable part is the *response to the “initial condition”*

$$\boldsymbol{u} = \begin{pmatrix} X_{t-1} \\ X_{t-2} \\ \ldots \\ X_{t-p} \end{pmatrix}. \tag{25}$$

With this notation, $\hat{S}_X = \boldsymbol{u}^T \widehat{\boldsymbol{\phi}}_X$ is the estimated predictable response of X_t. The estimated predictable response of Y_t can be defined analogously and denoted $\hat{S}_Y$. The initial conditions $\boldsymbol{u}$ may be drawn from the stationary distribution of X_t, the stationary distribution Y_t, or some mixture of the two. In fact, the covariance matrix $\boldsymbol{\Sigma}_{\text{HM}}$ defined in Eq. (A40) is proportional to the “harmonic mean” of the sample time-lagged covariance matrices for X_t and Y_t. Assuming that the initial condition $\boldsymbol{u}$ is drawn from a distribution with covariance matrix $\boldsymbol{\Sigma}_{\text{HM}}$ independently of $\boldsymbol{u}$, then

$$\operatorname{var}\left[\hat{S}_X\right] = \operatorname{var}\left[\boldsymbol{u}^T \widehat{\boldsymbol{\phi}}_X\right] = \boldsymbol{\phi}_X^T \boldsymbol{\Sigma}_{\text{HM}} \boldsymbol{\phi}_X. \tag{26}$$

If the initial condition for X_t and Y_t is contrived to be the same $\boldsymbol{u}$, then

$$\begin{aligned} \operatorname{var}\left[\hat{S}_X - \hat{S}_Y\right] &= \operatorname{var}\left[\boldsymbol{u}^T\left(\boldsymbol{\phi}_X - \boldsymbol{\phi}_Y\right)\right] \\ &= \left(\boldsymbol{\phi}_X - \boldsymbol{\phi}_Y\right)^T \boldsymbol{\Sigma}_{\text{HM}} \left(\boldsymbol{\phi}_X - \boldsymbol{\phi}_Y\right). \end{aligned} \tag{27}$$

Comparing this expression to $F_{\phi|\sigma}$ in Table 1 suggests that $F_{\phi|\sigma}$ is a kind of signal-to-noise ratio, where the “signal” is a *difference* in predictable responses for the same initial condition.

This difference in predictable responses, and hence $F_{\phi|\sigma}$, can be related to the difference in R-squares of the individual time series. To see this, expand Eq. (27) as

$$\operatorname{var}[S_X - S_Y] = \operatorname{var}[S_X] + \operatorname{var}[S_Y] - 2\boldsymbol{\phi}_X^T \boldsymbol{\Sigma}_{\text{HM}} \boldsymbol{\phi}_Y. \tag{28}$$

The Cauchy–Schwarz inequality implies that

$$\boldsymbol{\phi}_X^T \boldsymbol{\Sigma}_{\text{HM}} \boldsymbol{\phi}_Y \leq \sqrt{\operatorname{var}[S_X] \operatorname{var}[S_Y]}. \tag{29}$$

Hence, the above two expressions imply that

$$\operatorname{var}[S_X - S_Y] \geq \left(\sqrt{\operatorname{var}[S_X]} - \sqrt{\operatorname{var}[S_Y]}\right)^2. \tag{30}$$

In the special case of equal noise variances, the above inequality becomes

$$\frac{\mathrm{var}[S_X - S_Y]}{\sigma^2} \geq \left(\sqrt{\mathrm{SNR}_X} - \sqrt{\mathrm{SNR}_Y}\right)^2, \tag{31}$$

where SNR_X and SNR_Y are the signal-to-noise ratios of the two AR models:

$$\mathrm{SNR}_X = \frac{\mathrm{var}[S_X]}{\sigma_X^2} \quad \text{and} \quad \mathrm{SNR}_X = \frac{\mathrm{var}[S_Y]}{\sigma_Y^2}. \tag{32}$$

Note that $pF_{\phi|\sigma}$ is an estimate of the left-hand side of Eq. (31). Recall that R-square is defined as one minus the ratio of the noise variance to the total variance:

$$R_X^2 = 1 - \frac{\mathrm{var}[\epsilon_t^X]}{\mathrm{var}[X_t]} = \frac{\mathrm{SNR}_X}{\mathrm{SNR}_X + 1}. \tag{33}$$

Because R-square and signal-to-noise ratio are one-to-one, inequality Eq. (31) implies the following: if the noise variances are identical, then a large difference in predictabilities (i.e., R-squares) necessarily implies a tendency for $F_{\phi|\sigma}$ to be large. This suggests that it may be informative to compare R-squares. We use the sample estimate

$$\hat{R}_X^2 = 1 - \frac{\nu_X \hat{\sigma}_X^2}{\sum_{t=p+1}^{N_X} \left(X_t - \overline{X}\right)^2}$$
$$\text{where} \quad \overline{X} = \frac{1}{N_X - p} \sum_{t=p+1}^{N_X} X_t, \tag{34}$$

which is always between 0 and 1.

It should be recognized that a difference in R-square is a sufficient, but not necessary, condition for a difference in predictable responses. This fact can be seen from the fact that R-square for an AR(p) model is

$$R_X^2 = \phi_1^X \rho_X(1) + \ldots + \phi_p^X \rho_X(p). \tag{35}$$

In particular, two processes may have the same R-square because the *combination* of ϕs yields the same R-square, but the *specific values* of the ϕs may differ. Although a difference in R-squares is not a perfect indicator of $F_{\phi|\sigma}$, it is at least a sufficient condition and therefore worth examination.

On the other hand, if the R-squares are the same, a difference in process still could be detected and should be explained. Simply identifying differences in ϕs would be unsatisfying because those differences have a complicated relation to the statistical characteristics of the process when $p > 1$. Accordingly, we propose the following approach, which despite its limitations may still be insightful. Our basic idea is to *choose initial conditions that maximize the mean square difference in predictable responses*. To be most useful, we want this choice of initial condition to account for a *multimodel* comparison over M models. Accordingly, we define the mean square difference between predictable responses as

$$\Gamma[\boldsymbol{u}] = \sum_{m=1}^{M} \sum_{m'=1}^{M} \| S_m(\boldsymbol{u}) - S_{m'}(\boldsymbol{u}) \|^2$$
$$= \sum_{m=1}^{M} \sum_{m'=1}^{M} \left\| \boldsymbol{u}^T \left(\boldsymbol{\phi}_m - \boldsymbol{\phi}_{m'}\right) \right\|^2 = \boldsymbol{u}^T \mathbf{A} \boldsymbol{u}, \tag{36}$$

where

$$\mathbf{A} = \sum_{m=1}^{M} \sum_{m'=1}^{M} \left(\boldsymbol{\phi}_m - \boldsymbol{\phi}_{m'}\right) \left(\boldsymbol{\phi}_m - \boldsymbol{\phi}_{m'}\right)^T. \tag{37}$$

Our goal is to choose the initial condition $\boldsymbol{u}$ that maximizes $\Gamma[\boldsymbol{u}]$. The initial condition $\boldsymbol{u}$ must be constrained in some way, otherwise $\Gamma[\boldsymbol{u}]$ is unbounded and there is no maximum. If the initial conditions are drawn from a distribution with covariance matrix $\boldsymbol{\Sigma}_M$, then an appropriate constraint is to fix the associated Mahalanobis distance:

$$\boldsymbol{u}^T \boldsymbol{\Sigma}_M^{-1} \boldsymbol{u} = 1. \tag{38}$$

The problem is now to maximize $\boldsymbol{u}^T \mathbf{A} \boldsymbol{u}$ subject to the constraint $\boldsymbol{u}^T \boldsymbol{\Sigma}_M^{-1} \boldsymbol{u} = 1$. This is a standard optimization problem that is merely a generalization of principal component analysis (also called empirical orthogonal function analysis). The solution is obtained by solving the generalized eigenvalue problem

$$\mathbf{A}\boldsymbol{u} = \lambda \boldsymbol{\Sigma}_M^{-1} \boldsymbol{u}. \tag{39}$$

This eigenvalue problem yields p eigenvalues and p eigenvectors. The eigenvalues can be ordered from largest to smallest, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$, and the corresponding eigenvalues can be denoted $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_p$. The first eigenvector $\boldsymbol{u}_1$ gives the initial condition that maximizes the sum square difference in predictable responses; the second eigenvector $\boldsymbol{u}_2$ gives the initial condition that maximizes Γ subject to the condition that $\boldsymbol{u}_2^T \boldsymbol{\Sigma}_M^{-1} \boldsymbol{u}_1 = 0$; and so on. The eigenvalues $\lambda_1, \ldots, \lambda_p$ give the corresponding values of Γ. The eigenvectors can be collected into the $p \times p$ matrix

$$\mathbf{U} = \begin{bmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 & \ldots & \boldsymbol{u}_p \end{bmatrix}. \tag{40}$$

Because the matrices $\mathbf{A}$ and $\boldsymbol{\Sigma}_M$ are symmetric, the eigenvectors can be chosen to satisfy the orthogonality property

$$\mathbf{U}^T \boldsymbol{\Sigma}_M^{-1} \mathbf{U} = \mathbf{I} \quad \Rightarrow \quad \boldsymbol{\Sigma}_M = \mathbf{U}\mathbf{U}^T. \tag{41}$$

Summing Γ over all eigenvectors gives

$$\begin{aligned}\sum_{k=1}^{p}\Gamma[\boldsymbol{u}_k] &= \sum_{k=1}^{p}\boldsymbol{u}_k^T\mathbf{A}\boldsymbol{u}_k = \operatorname{tr}\left[\mathbf{A}\sum_{k=1}^{p}\boldsymbol{u}_k\boldsymbol{u}_k^T\right]\\ &= \operatorname{tr}\left[\mathbf{A}\mathbf{U}\mathbf{U}^T\right] = \operatorname{tr}\left[\mathbf{A}\boldsymbol{\Sigma}_M\right]\\ &= \sum_{m=1}^{M}\sum_{m'=1}^{M}\left(\boldsymbol{\phi}_m-\boldsymbol{\phi}_{m'}\right)^T\\ &\quad \boldsymbol{\Sigma}_M\left(\boldsymbol{\phi}_m-\boldsymbol{\phi}_{m'}\right). \qquad (42)\end{aligned}$$

Note the similarity of the final expression to Eq. (27). This similarity implies that if only two models are compared and $\boldsymbol{\Sigma}_M = \boldsymbol{\Sigma}_{\mathrm{HM}}$, then the solution exactly matches the variance of signal differences (27). Unfortunately, $\boldsymbol{\Sigma}_{\mathrm{HM}}$ and the noise variance $\hat{\sigma}^2$ depend on (m, m'), so it is difficult to generalize the optimal initial condition to explain all pair-wise values of $F_{\phi|\sigma}$. We suggest using a covariance matrix that is the harmonic mean across all M time series:

$$\boldsymbol{\Sigma}_M = M\left(\widehat{\boldsymbol{\Sigma}}_1^{-1}+\widehat{\boldsymbol{\Sigma}}_2^{-1}+\ldots+\widehat{\boldsymbol{\Sigma}}_M^{-1}\right)^{-1}, \qquad (43)$$

where $\widehat{\boldsymbol{\Sigma}}_m$ is the sample covariance matrix of the initial condition Eq. (25) for the mth time series. Finally, we note that

$$\Gamma_{\mathrm{sum}} = \sum_{k=1}^{p}\Gamma[\boldsymbol{u}_k] = \lambda_1+\lambda_2+\ldots+\lambda_p. \qquad (44)$$

Thus, $\lambda_k/\Gamma_{\mathrm{sum}}$ is the fraction of sum total variance of signal differences explained by the kth eigenvector. Conceptually, each eigenvector $\boldsymbol{u}_k$ can be interpreted as an initial condition that has been optimized to separate predictable responses.

5 Example: diagnosing differences in AMOC simulations

To illustrate the proposed method, we apply it to an index of the Atlantic Meridional Overturning Circulation (AMOC). This variable is chosen because it is considered to be a major source of decadal variability and predictability (Buckley and Marshall, 2016; Zhang et al., 2019). However, the AMOC is not observed directly with sufficient frequency and consistency to constrain its variability on decadal timescales. As a result, there has been a heavy reliance on coupled atmosphere–ocean models to characterize AMOC variability and predictability. The question arises as to whether simulations of the AMOC by different models can be distinguished and, if so, how they differ.

The data used in this study come from pre-industrial control runs from Phase 5 of the Coupled Model Intercomparison Project (CMIP5; Taylor et al., 2010). These control simulations lack year-to-year changes in forcing and thereby permit a focus on internal variability without confounding effects due to anthropogenic climate change. We consider only models that have pre-industrial control simulations spanning at least 500 years and contain meridional overturning circulation as an output variable. Only 10 models meet these criteria. An AMOC index is defined as the annual mean of the maximum meridional overturning streamfunction at 40° N in the Atlantic. A third-order polynomial over the 500 years is removed to eliminate climate drift. The AMOC index from the 10 models is shown in Fig. 1. Based on visual comparisons, one might perceive differences in amplitude (e.g., the MPI models tend to have larger amplitudes than other models) and differences in the degree of persistence (e.g., high-frequency variability is more evident in INMCM4 than in other models), but whether these differences are statistically significant remains to be established.

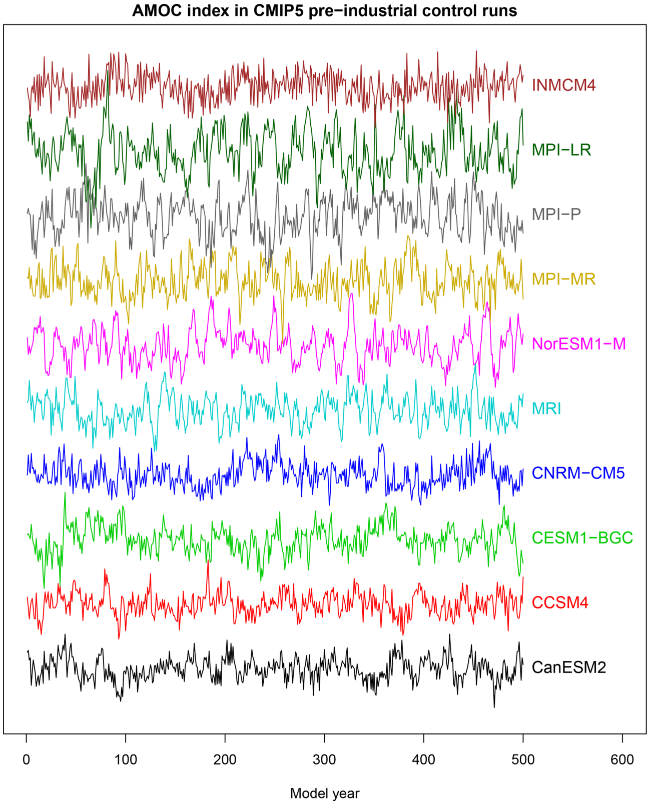


Figure 1. AMOC index simulated by 10 CMIP5 models under pre-industrial conditions. Each time series is offset by a constant with no re-scaling. The AMOC index is defined as the maximum annual mean meridional overturning streamfunction at 40° N in the Atlantic.

To compare autoregressive processes, it is necessary to select the order of the autoregressive model. As discussed in Sect. 3, we use Eq. (22), which for $\nu = 1.0$ gives $p^* = 5$. One check on this choice is whether the residuals are white noise. We find that AR(5) is adequate for all models except MRI, in the sense that the residuals reveal no serious departures from white noise according to visual inspection and by the Ljung–Box test (Ljung and Box, 1978). Another check is

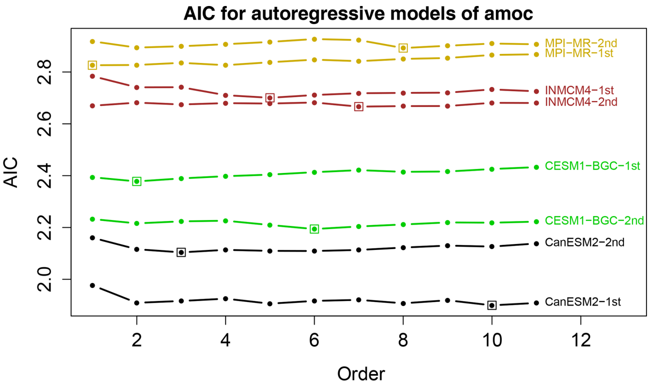


Figure 2. Akaike's information criterion (AIC) as a function of autoregressive model order for AMOC time series from selected CMIP5 models. AIC is computed separately from the first and second 250 years of the data. A box identifies the minimum AIC up to order 11. The actual AIC values are shown – AIC values have not been offset. Models have been selected primarily to avoid line crossings.

to compute Akaike's information criterion (AIC; Box et al., 2008) for the two halves of the data (i.e., 250 years). Some representative examples are shown in Fig. 2. As can be seen, AIC is a relatively flat function of model order, hence small sampling variations can lead to large changes in order selection. Indeed, the order selected by AIC is sensitive to which half of the data is used. Nevertheless, because AIC is nearly a flat function of order, virtually any choice of order beyond AR(2) can be defended. The highest order selected by AIC is $p = 11$. While we have performed our test for AR(11), this order would be a misleading illustration of the method because one might assume that 11 lags are *necessary* to identify model differences. Thus, in the results that follow, we choose AR(5) for all cases and bear in mind that comparisons with MRI may be affected by model inadequacy.

The choice of AR(5) means that all information relevant to deciding differences in stationary processes is contained in the first six values of the sample auto-covariance function $\hat{c}(0), \ldots, \hat{c}(5)$. The sample auto-covariance function for each AMOC index is shown in Fig. 3. Recall that the zero-lag auto-covariance $\hat{c}(0)$ is the sample variance. The figure suggests that the time series have different variances and different decay rates, but it is unclear whether these differences are significant. Note that a standard difference-in-variance test cannot be performed here because the time series are serially correlated. One might try to modify the standard F-test by adjusting the degrees of freedom, as is sometimes advocated in the t-test (Zwiers and von Storch, 1995), but the adjustment depends on the autocorrelation function that we are trying to compare. An alternative approach is to pre-whiten the data based on the AR fit and then test differences in variance, but this is exactly equivalent to our test based on F_σ.

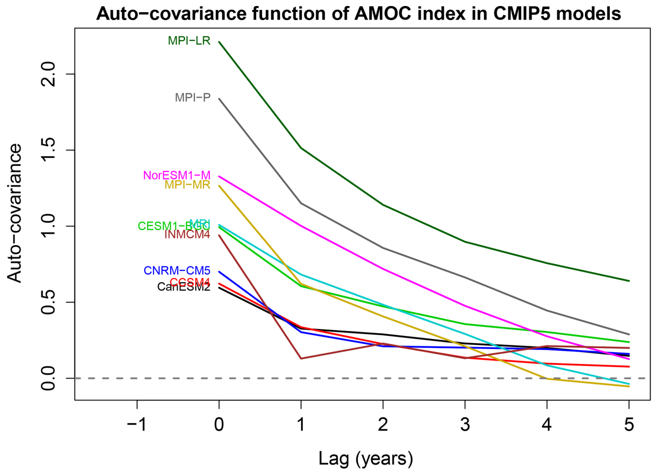


Figure 3. Auto-covariance function of the AMOC time series from each CMIP5 model, as estimated from Eq. (6) using the first 250 years of data.

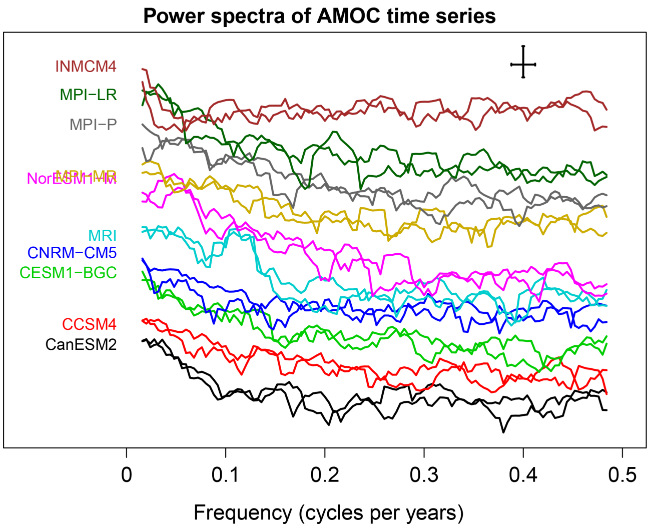
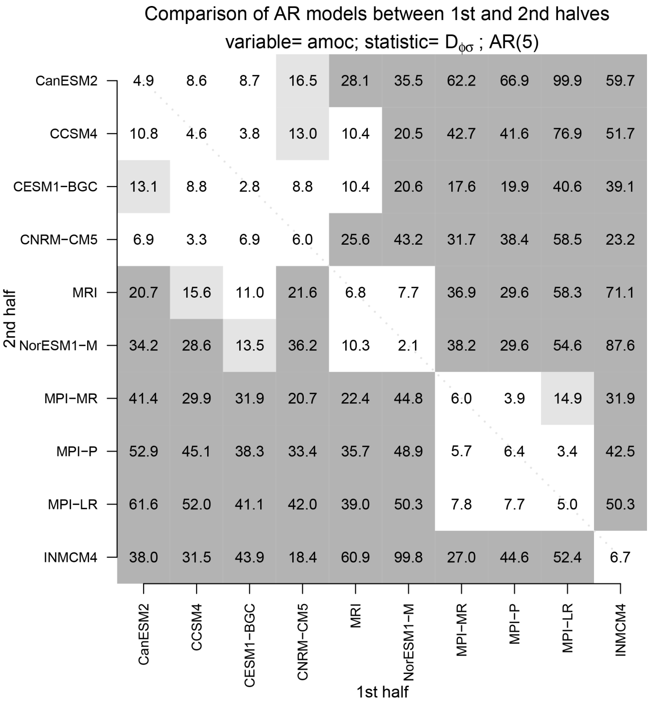


Figure 4. Power spectra of AMOC time series from CMIP5 models. Spectra are estimated from the first and second 250 years of each time series using Daniell's estimator. The 95 % confidence interval and bandwidth are indicated by the error bars in the top right corner. The power spectra have been offset by a multiplicative constant to reduce overlap (the y axis is log scale). The longest resolved period is 60 years.

An alternative approach to comparing stationary processes is to compare power spectra. Power spectra of the AMOC time series are shown in Fig. 4. The spectra have been offset by a multiplicative constant to reduce overlap (otherwise the different curves would obscure each other). While many differences can be seen, the question is whether those differences are significant after accounting for sampling uncertainties. The two spectral tests discussed in Sect. 2 find relatively few differences between CMIP5 models, although they do indicate that time series from INMCM4 and MPI differ from


Comparison of AR models between 1st and 2nd halves
variable= amoc; statistic= D$_{\phi\sigma}$; AR(5)
2nd half
1st half

2nd half \ 1st half	CanESM2	CCSM4	CESM1-BGC	CNRM-CM5	MRI	NorESM1-M	MPI-MR	MPI-P	MPI-LR	INMCM4
CanESM2	4.9	8.6	8.7	16.5	28.1	35.5	62.2	66.9	99.9	59.7
CCSM4	10.8	4.6	3.8	13.0	10.4	20.5	42.7	41.6	76.9	51.7
CESM1-BGC	13.1	8.8	2.8	8.8	10.4	20.6	17.6	19.9	40.6	39.1
CNRM-CM5	6.9	3.3	6.9	6.0	25.6	43.2	31.7	38.4	58.5	23.2
MRI	20.7	15.6	11.0	21.6	6.8	7.7	36.9	29.6	58.3	71.1
NorESM1-M	34.2	28.6	13.5	36.2	10.3	2.1	38.2	29.6	54.6	87.6
MPI-MR	41.4	29.9	31.9	20.7	22.4	44.8	6.0	3.9	14.9	31.9
MPI-P	52.9	45.1	38.3	33.4	35.7	48.9	5.7	6.4	3.4	42.5
MPI-LR	61.6	52.0	41.1	42.0	39.0	50.3	7.8	7.7	5.0	50.3
INMCM4	38.0	31.5	43.9	18.4	60.9	99.8	27.0	44.6	52.4	6.7

Figure 5. A measure of the "distance" between AMOC time series between the first and second halves of CMIP5 pre-industrial control simulations. The distance is measured by the bias-corrected deviance statistic $D_{\phi,\sigma}$ using an fifth-order AR model. The light and dark gray shadings show values exceeding the 5 % and 1 % significance levels, respectively (the threshold values are 12.7 and 17.0, respectively).

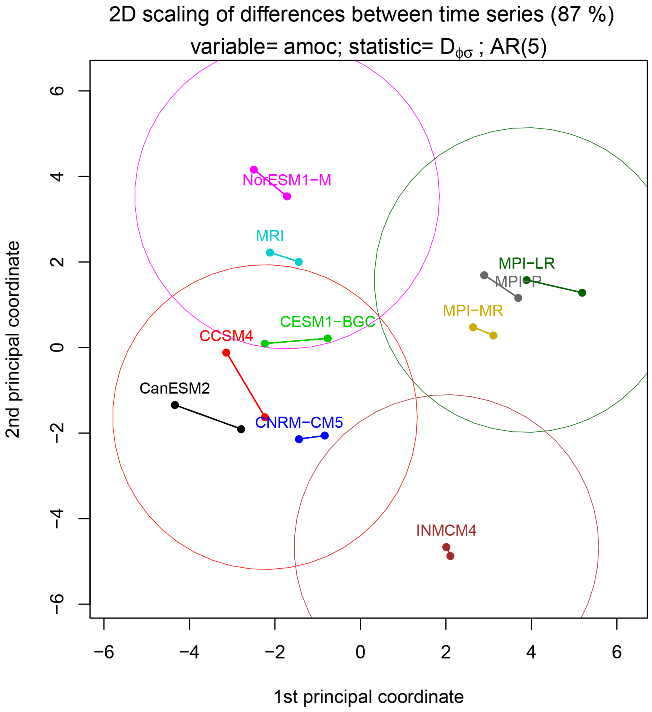


Figure 6. A set of points in a two-dimensional Cartesian plane whose pair-wise Euclidean distances most closely approximate the pair-wise distances between autoregressive processes of AMOC time series. The difference between AR processes is measured by the bias-corrected deviance statistic $D_{\phi,\sigma}$ and the points are identified using multidimensional scaling techniques. There are 20 points corresponding to 10 CMIP5 models, each model time series being split in half. Time series from the same model have the same color and are joined by a line segment. Circles around selected points enclose models whose time series are statistically indistinguishable from that of the center model at the 5 % significance level.

those of other CMIP5 models (not shown). Presumably, these differences arise from the fact that the INMCM4 time series is closer to white noise and that the MPI time series have larger total variance than those from other CMIP5 models.

To illustrate our proposed method, we perform the following analysis. First, the AMOC index from each CMIP5 model is split into equal halves, each 250 years long. Then, each time series from the first half is compared to each time series in the second half. Some of these comparisons will involve time series from the same CMIP5 model. Our expectation is that no difference should be detected when time series from two different halves of the same CMIP5 model are compared. To summarize the comparisons, we show in Fig. 5 a matrix of the bias-corrected deviance statistic $D_{\phi,\sigma}$ for every possible model comparison. This statistic is a measure of the "distance" between stationary process. The two shades of gray indicate a significant difference in AR process at the 5 % or 1 % level. Values along the diagonal correspond to comparisons of AMOC time series from the same CMIP5 model. No significant difference is detected when time series come from the same model. In contrast, significant differences are detected in most of the cases when the time series come from different CMIP5 models. Interestingly, models from the same institution tend to be indistinguishable from each other. For instance, the MPI models are mostly indistinguishable from each other, and the CCSM and CESM-BGC models (from the National Center for Atmospheric Research, NCAR) are mostly indistinguishable from each other. We say "mostly" because the difference depends on which half of the simulation is used in the comparison. For instance, a difference is detected when comparing the first half of the CESM1-BGC time series to the second half to the CCSM4 time series, but not vice versa. The models have been ordered to make natural clusters easy to identify in this figure. Aside from a few exceptions, the pattern of shading is the same for AR(2) and AR(10) models (not shown), demonstrating that our conclusion regarding significant differences in AR process is not sensitive to model order. Also, the pattern of shading is similar if the comparison is based on only from the first half of the simulations, or only from the second half of the simulations (not shown), indicating that our results are not sensitive to sampling errors.

To visualize natural clusters more readily, we identify a set of points in a cartesian plane whose pair-wise distances match the above distances as closely as possible. These points can be identified using a procedure called *multidimensional scaling* (Izenman, 2013). The procedure is to compute the deviance statistic between every possible pair of time series. Because there are 10 CMIP models and time series from each model is split in half, there are 20 time series being compared. Thus, the deviance statistic for every possible pair can be summarized in a 20×20 distance matrix. From this matrix, multidimensional scaling can find the set of points in a 20-dimensional space that has this distance matrix. Moreover, it can find the points in a two-dimensional space whose distance matrix most closely approximates the original distance matrix in some objective sense. In our case, 87 % of original distance matrix can be represented by two-dimensional points. These points are shown in Fig. 6. Although 87 % of the distance matrix is represented in this figure, isolated points may have relatively large discrepancies. The average discrepancy is about 2 units, with the largest discrepancy being about 3.6 units between MRI and MPI-LR. An attractive feature of this representation is that the decision rule for statistical significance, $D_{\phi,\sigma} > D_{\text{crit}}$, is approximately equivalent to drawing a circle of radius $\sqrt{D_{\text{crit}}}$ around a point and identifying all the points that lie outside that circle. This geometric procedure would be exact in a 20-dimensional space, but is only approximate in the 2-dimensional space shown in Fig. 6. The figure suggests that the MPI models and INMCM4 form their own natural clusters. For other models, a particular choice of clusters is shown in the figure, but this is merely an example, and different clusters with different groupings could be justified. Note that all line segments connecting results from the same CMIP model are shorter than the circle radius, indicating no significant difference between time series from the same CMIP model.

It is interesting to relate the above differences to the auto-covariance functions shown in Fig. 3. It is likely that INMCM4 differs from other models because its auto-covariance function decays most quickly to zero. The MPI models are distinguished from the others by their large variances. Interestingly, note that the auto-covariance function for NorESM1-M is *intermediate* between that of two MPI models, yet the test indicates that NorESM1-M differs from the MPI models. Presumably, the MPI models are indistinguishable because their auto-covariance functions have the same shape, including a kink at lag 1, whereas the NorESM1-M model has no kink at lag 1. This example illustrates that the test does not behave like a mean square difference between auto-covariance functions, which would cluster NorESM1-M with the MPI models.

It is worth clarifying how the above clustering technique differs from those in previous studies. Piccolo (1990) proposed a distance measure based on the Euclidean norm of the difference in AR parameters, namely

$$d(X,Y) = \left\{ \sum_{\tau=1}^{\infty} \left(\phi_\tau^X - \phi_\tau^Y \right)^2 \right\}^{1/2}. \qquad (45)$$

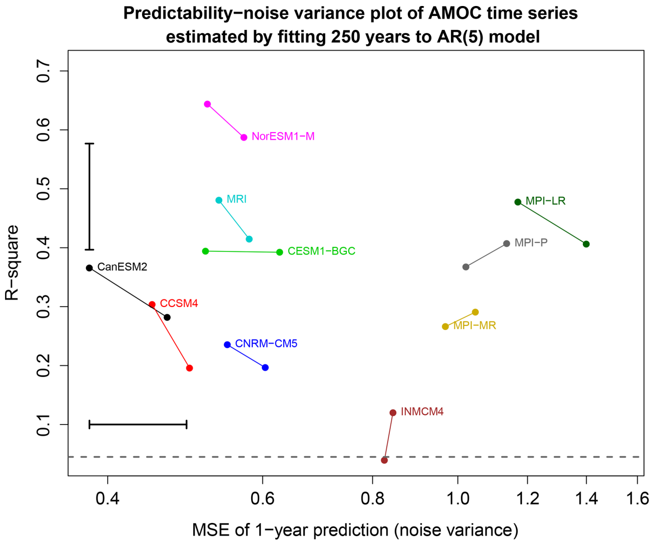


Figure 7. Noise variance versus the R-square of AR(5) models estimated from 250-year segments of AMOC time series from 10 CMIP5 models. Estimates from the same CMIP5 model have the same color and are joined by a line segment. The error bar in the bottom left corner shows the critical distance for a significant difference in noise variance at the 5 % level. The x axis is on a log scale so that equal variance ratios correspond to equal linear distances. The error bar in the upper left shows a 95 % confidence interval for R-square using the method of Lee (1971) and computed from the `MBESS` package in R. The dashed line is the 5 % significance threshold for the R-square.

In contrast, our hypothesis test uses a Mahalanobis norm for measuring differences in AR parameters, where the covariance matrix is based on the sample covariance matrices of the time-lagged data (see Eq. A44). While the Euclidean norm Eq. (45) does have some attractive properties as discussed in Piccolo (1990), it is inconsistent with the corresponding hypothesis test for differences in AR parameters. As a result, the resulting cluster may emphasize differences with large Euclidean norms that are insignificant, or may downplay differences in small Euclidean norms that are significant. In contrast, the distance measure used in our study is consistent with a rigorous hypothesis test. Maharaj (2000) proposes a classification procedure that is consistent with hypothesis testing, but that hypothesis test does not account for differences in noise variances.

Having identified differences between stationary processes, it is of interest to relate those differences to standard properties of the time series. Recall that our measure of the difference between stationary processes $D_{\phi,\sigma}$ is the sum of two other measures, namely D_σ and $D_{\phi|\sigma}$. Measure D_σ depends only on the noise variances, that is, it depends on the mean square error of a 1-year prediction. In contrast, $D_{\phi|\sigma}$ measures the difference in predictable responses of the process. As discussed in Sect. 4, we suggest examining differences in R-square. A graph of the noise variance plotted

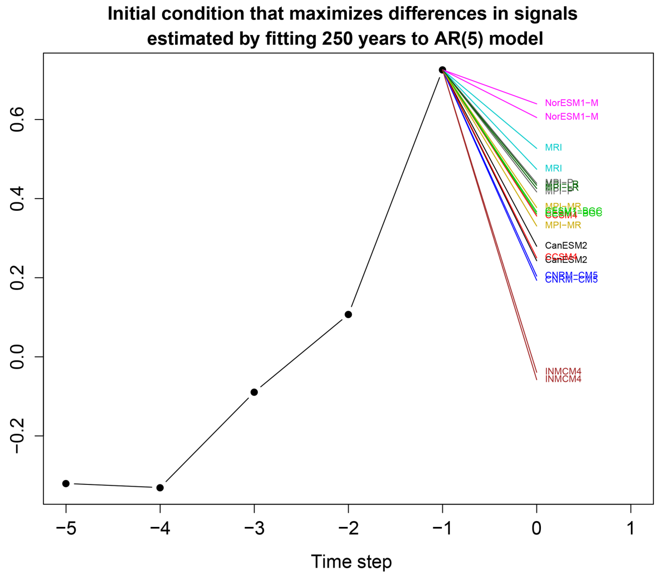


Figure 8. Predictions from each estimated AR(5) model using the optimal initial condition derived from Eq. (39). The optimal initial condition is the five black dots joined by lines, and the resulting predictions are the colored lines. Each CMIP model has two predictions corresponding to the two AR(5) models estimated from two non-overlapping 250-year time series from the same CMIP model.

against R-square of each autoregressive model is shown in Fig. 7. The error bars show the critical distance for a significance difference in noise variance (lower left) and for a difference in R-square (upper right). First note that the projections of line segments onto the x axis or y axis are shorter than the respective error bars, indicating that differences in noise and differences in predictability are insignificant when estimated from the *same* CMIP5 model. Second, note that the relative positions of the dots are similar to those in Fig. 6. Thus, the noise variance and R-square appear to be approximate "principal coordinates" for distinguishing univariate autoregressive processes. Third, using noise variances alone, the MPI models would be grouped together, then INMCM4 would be grouped by itself, while at the bottom end CanESM2 and CCSM4 would be grouped together. These groupings are consistent with the clusters identified above using the full distance measure $D_{\phi,\sigma}$, suggesting that differences in noise variances explain the major differences between stationary processes. Fourth, the AR models have R-square values mostly between 0.25 and 0.5. Time series from INMCM4 have the smallest R-square values while time series from NorESM1 have the largest R-square values.

An alternative approach to describing differences in AR parameters is to show differences in response to the same initial condition. We use the optimization method discussed in Sect. 4 to find the initial condition that maximizes the sum square difference in responses. The result is shown in Fig. 8. Essentially, the models separate by predictability – the order of the models from top to bottom closely tracks the order of the models based on R-square seen in Fig. 7. For this initial condition, INMCM4 damps nearly to zero in one time step, whereas NorESM1-M decays the slowest among the models, consistent with expectations from R-square. This initial condition explains 82 % of the differences in response. Because optimal initial conditions form a complete set, an arbitrary initial condition can be represented as a linear combination of optimal initial conditions. If the AR models were used to predict an observational time series with covariance matrix $\mathbf{\Sigma}_M$, then most differences between model predictions could be explained by the differences in response to a single optimal initial condition.

We end with a brief summary of our exploration of the χ^2 statistic proposed by Lund et al. (2009). As mentioned earlier, the result is sensitive to the choice of K. However, instead of summing based on the cutoff parameter K, we summed over all possible lags, and set all sample autocovariances beyond lag p to zero. Using this approach, we find that the resulting χ^2 statistic gives results very similar to ours for $p = 5$, including clustering the MPI models with each other, and separating them from NorEMS1-M. We have not investigated this procedure sufficiently to propose it as a general rule but mention it to suggest the possibility that some alternative rule for computing the sum Eq. (9) may yield reasonable results.

6 Conclusions

This paper examined tests for differences in stationary processes and proposed a new approach to characterizing those differences. The basic idea is to fit each time series to an autoregressive model and then test for differences in parameters. The likelihood ratio test for this comparison was derived in Maharaj (2000) and Grant and Quinn (2017). We have modified the test to correct for certain biases and to include a test for differences in noise variance. The latter test is of major interest in climate applications and leads to considerable simplifications in estimation and interpretation. Furthermore, the proposed test is applicable to small samples. In addition, we propose new approaches to interpreting and visualizing differences in stationary processes. The procedure was illustrated on AMOC time series from pre-industrial control simulations of 10 models in the CMIP5 archive. Based on time series 250 years in length, the procedure was able to distinguish about four clusters of models, where time series from the same CMIP5 model are grouped together in the same cluster. These clusters are identified easily using a multidimensional technique. The clusters obtained by this method were not sensitive to the order of the autoregressive model, although the number of significant differences decreases with AR order due to the larger number of parameters being estimated. Further analysis shows that these clusters can be explained largely by differences in 1-year prediction errors in the AR models and differences in R-square.

The proposed method can be used to compare any stationary time series that are well fit by an autoregressive model, which includes most climate time series. Thus, this method could be used to decide whether two climate models generate the same temporal variability. A natural question is whether this approach can be generalized to compare multivariate time series. This generalization will be developed in Part 2 of this paper. The method also could be used to compare model simulations to observations, provided that the stationarity assumption is satisfied. If non-stationarity is strong, then this method would need to be modified to account for such non-stationarity, such as adding exogenous terms to the AR model. The likelihood-ratio framework can accommodate such extensions and will be developed in Part 3 of this paper.

Appendix A: Derivation of the test

In this Appendix, we derive a test for equality of parameters of autoregressive models based on the likelihood ratio test. The derivation is similar to that in Maharaj (2000) and Grant and Quinn (2017), except modified to test for differences in noise variance and to correct for known biases in maximum likelihood estimates. In addition, we show how the bias-corrected likelihood ratio can be partitioned into two independent ratios and derive the sampling distributions for each.

We consider only the *conditional* likelihood, which is the likelihood function conditioned on the first p values of the process. The conditional likelihood approach is reasonable for large sample sizes and has the advantage that the estimates can be obtained straightforwardly from the method of least squares. Suppose we have a time series of length N_X for X_t. Then the conditional likelihood function for X_t is

$$L_X = (2\pi\sigma_X)^{-(N_X-p)/2} \exp\left[-\frac{\sum_{t=p+1}^{N_X}\left(X_t - \phi_1^X X_{t-1} - \ldots - \phi_p^X X_{t-p} - \gamma^X\right)^2}{2\sigma_X^2} \right] \tag{A1}$$

(see Eq. A7.4.2b of Box et al., 2008). The maximum likelihood estimates of the parameters, denoted by $\hat{\phi}_1^X, \ldots, \hat{\phi}_p^X, \hat{\gamma}^X$, are obtained from the method of least squares. The maximum likelihood estimate of σ_X^2 is

$$\overline{\sigma}_X^2 = \frac{\sum_{t=p+1}^{N_X}\left(X_t - \hat{\phi}_1^X X_{t-1} - \ldots - \hat{\phi}_p^X X_{t-p} - \hat{\gamma}^X\right)^2}{N_X - p}. \tag{A2}$$

Substituting this into the likelihood function Eq. (A1) and then taking -2 times the logarithm of the likelihood function gives

$$-2\log \overline{L}_X = (N_X - p)\left(\log \overline{\sigma}_X^2 + \log 2\pi + 1\right). \tag{A3}$$

Similarly, suppose we have a time series of length N_Y for Y_t. Then, the corresponding likelihood function is

$$L_Y = (2\pi\sigma_Y)^{-(N_Y-p)/2} \exp\left[-\frac{\sum_{t=p+1}^{N_Y}\left(Y_t - \phi_1^Y Y_{t-1} - \ldots - \phi_p^Y Y_{t-p} - \gamma^Y\right)^2}{2\sigma_Y^2} \right],$$

the maximum likelihood estimate of σ_Y^2 is

$$\overline{\sigma}_Y^2 = \frac{\sum_{t=p+1}^{N_Y}\left(Y_t - \hat{\phi}_1^Y Y_{t-1} - \ldots - \hat{\phi}_p^Y Y_{t-p} - \hat{\gamma}^Y\right)^2}{N_Y - p}, \tag{A4}$$

and -2 times the logarithm of the likelihood function is

$$-2\log \overline{L}_Y = (N_Y - p)\left(\log \overline{\sigma}_Y^2 + \log 2\pi + 1\right). \tag{A5}$$

Because X_t and Y_t are independent, the likelihood function for all the data is the product $L_X L_Y$.

Under hypothesis H_0, the likelihood function has the same form as $L_X L_Y$, except there is only a single set of autoregressive parameters $\phi_1, \ldots, \phi_p$ and a single noise variance σ. The corresponding likelihood function is therefore

$$L_{\sigma,\phi} = (2\pi\sigma)^{-(N_X-p)/2}(2\pi\sigma)^{-(N_Y-p)/2} \exp\left[\frac{\begin{array}{c} -\sum_{t=p+1}^{N_X}(X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} - \gamma^X)^2 \\ -\sum_{t=p+1}^{N_Y}(Y_t - \phi_1 Y_{t-1} - \ldots - \phi_p Y_{t-p} - \gamma^Y)^2 \end{array}}{2\sigma^2} \right]. \tag{A6}$$

The maximum likelihood estimates of $\phi_1, \ldots, \phi_p$, denoted $\hat{\phi}_1, \ldots, \hat{\phi}_p$, are obtained from the least squares estimates of the *pooled* sample. Again, these estimates are obtained easily by the method of least squares. The maximum likelihood estimate of the common variance σ^2 is

$$\overline{\overline{\sigma}}^2 = \frac{\begin{array}{c} \sum_{t=p+1}^{N_X}\left(X_t - \hat{\phi}_1 X_{t-1} - \ldots - \hat{\phi}_p X_{t-p} - \hat{\gamma}^X\right)^2 \\ +\sum_{t=p+1}^{N_Y}\left(Y_t - \hat{\phi}_1 Y_{t-1} - \ldots - \hat{\phi}_p Y_{t-p} - \hat{\gamma}^Y\right)^2 \end{array}}{N_X + N_Y - 2p}. \tag{A7}$$

The corresponding log-likelihood is

$$-2\log \overline{L}_{\sigma,\phi} = (N_X + N_Y - 2p)\left(\log \overline{\overline{\sigma}}^2 + \log 2\pi + 1\right). \tag{A8}$$

Finally, we compute the likelihood ratio, or equivalently, the difference in the log-likelihood functions. This difference (multiplied by -2) is called the *deviance statistic* and is

$$\begin{aligned} \overline{D}_{\sigma,\phi} =& 2\log \overline{L}_X + 2\log \overline{L}_Y - 2\log \overline{L}_{\sigma,\phi} \\ =& \log\left(\frac{\overline{\overline{\sigma}}^{2(N_X+N_Y-2p)}}{\overline{\sigma}_X^{2(N_X-P)}\overline{\sigma}_Y^{2(N_Y-p)}} \right). \end{aligned} \tag{A9}$$

It is well known that maximum likelihood estimates of variance are biased. Accordingly, we define *bias-corrected* deviance statistics before proceeding. This can be done by replacing sample sizes by degrees of freedom. Care must be exercised in such substitutions because replacing MLEs with unbiased versions will yield likelihoods that are no longer maximized, and therefore may yield deviance statistics that are negative. Our goal is to define a bias-corrected deviance statistic that is non-negative

To compute the degrees of freedom, note that the sample size of X_t is $N_X - p$, because the first p time steps are excluded from the conditional likelihood, and $p+1$ parameters are being estimated, so the degrees of freedom is the difference, $N_X - 2p - 1$. Similarly, the degrees of freedom for Y_t is $N_Y - 2p - 1$. Let the degrees of freedom be denoted

$$\nu_X = N_X - 2p - 1 \quad \text{and} \quad \nu_Y = N_Y - 2p - 1. \tag{A10}$$

Accordingly, an unbiased estimate of σ_X^2 is

$$\hat{\sigma}_X^2 = \frac{\sum_{t=p+1}^{N_X}\left(X_t - \hat{\phi}_1^X X_{t-1} - \ldots - \hat{\phi}_p^X X_{t-p} - \hat{\gamma}^X\right)^2}{\nu_X}, \tag{A11}$$

and an unbiased estimate of σ_Y^2 is

$$\hat{\sigma}_Y^2 = \frac{\sum_{t=p+1}^{N_Y}\left(Y_t - \hat{\phi}_1^Y Y_{t-1} - \ldots - \hat{\phi}_p^Y Y_{t-p} - \hat{\gamma}^Y\right)^2}{\nu_Y}. \tag{A12}$$

As will be shown below, the appropriate bias correction for $\overline{\overline{\sigma}}^2$ is

$$\hat{\hat{\sigma}}^2 = \overline{\overline{\sigma}}^2 \left(\frac{N_X + N_Y - 2p}{\nu_X + \nu_Y}\right). \tag{A13}$$

Also, it proves helpful to define the unbiased estimate of the common variance σ^2 as

$$\hat{\sigma}^2 = \frac{\nu_X \hat{\sigma}_X^2 + \nu_Y \hat{\sigma}_Y^2}{\nu_X + \nu_Y}. \tag{A14}$$

Then, the bias-corrected deviance statistic Eq. (A9) can be defined as

$$D_{\phi,\sigma} = D_\sigma + D_{\phi|\sigma}, \tag{A15}$$

where

$$D_\sigma = \log\left(\frac{\hat{\sigma}^{2\nu_X+2\nu_Y}}{\hat{\sigma}_X^{2\nu_X}\hat{\sigma}_Y^{2\nu_Y}}\right) \tag{A16}$$

$$D_{\phi|\sigma} = (\nu_X + \nu_Y)\log\left(\frac{\hat{\hat{\sigma}}^2}{\hat{\sigma}^2}\right). \tag{A17}$$

We now prove the D_σ and $D_{\phi|\sigma}$ are non-negative, independent, and have sampling distributions related to the F-distribution. To show that D_σ is non-negative, note that the ratio in Eq. (A16) is the weighted arithmetic mean over the weighted geometric mean of $\hat{\sigma}_X^2$ and $\hat{\sigma}_Y^2$. Consequently, we may invoke the AM-GM inequality to show that D_σ is non-negative and vanishes if and only if $\hat{\sigma}_X = \hat{\sigma}_Y$. In this sense, D_σ measures the "distance" between $\hat{\sigma}_X$ and $\hat{\sigma}_Y$. Incidentally, had we used the uncorrected likelihoods, the resulting deviance statistic would vanish at $\overline{\sigma}_X^2 = \overline{\sigma}_Y^2$, which would give a biased measure of deviance.

To prove the remaining properties of D_σ and $D_{\phi|\sigma}$, we adopt a vector notation that is better suited to the task. Accordingly, let the AR(p) model Eq. (11) be denoted

$$\boldsymbol{w}_X = \mathbf{Z}_X \boldsymbol{\phi}_X + \boldsymbol{j}\gamma_X + \boldsymbol{\epsilon}_X, \tag{A18}$$

where $\boldsymbol{w}_X$ is an $(N_X - p)$-dimensional vector, $\mathbf{Z}_X$ is a $(N_X - p) \times p$ matrix, $\boldsymbol{j}$ is a $(N_X - p)$-dimensional vector of ones, γ_x is a scalar, and the remaining terms have been defined previously.

$$\boldsymbol{w}_X = \begin{bmatrix} X_{p+1} \\ X_{p+2} \\ \vdots \\ X_{N_X} \end{bmatrix},$$

$$\mathbf{Z}_X = \begin{pmatrix} X_p & X_{p-1} & \ldots & X_1 \\ X_{p+1} & X_p & \ldots & X_2 \\ \vdots & \vdots & \ddots & \vdots \\ X_{N_X-1} & X_{N_X-2} & \ldots & X_{N_X-p} \end{pmatrix},$$

$$\boldsymbol{\phi}_X = \begin{pmatrix} \phi_1^X \\ \phi_2^X \\ \vdots \\ \phi_p^X \end{pmatrix}. \tag{A19}$$

Since H_0 does not restrict γ_X, it proves convenient to eliminate this parameter by projecting onto the complement of $\boldsymbol{j}$. Therefore, we multiply both sides of the equation by the projection matrix

$$\mathbf{H} = \mathbf{I} - \frac{1}{N_X - p}\boldsymbol{j}\boldsymbol{j}^T. \tag{A20}$$

This multiplication has the effect of eliminating the γ_X term and centering *each* column of $\boldsymbol{w}_X$ and $\mathbf{Z}_X$. Henceforth, we assume that each column of $\boldsymbol{w}_X$ and $\mathbf{Z}_X$ has been centered. One should remember that the degrees of freedom associated with estimating the noise variance σ_X^2 should be reduced by one to account for this pre-centering of the data. Similarly, the corresponding model for Y_t in Eq. (12) is written as

$$\boldsymbol{w}_Y = \mathbf{Z}_Y \boldsymbol{\phi}_Y + \boldsymbol{j}\gamma_Y + \boldsymbol{\epsilon}_Y, \tag{A21}$$

where $\boldsymbol{w}_Y$ is an $(N_Y - p)$-dimensional vector, and the remaining terms are analogous to those in Eq. (A18). As for X, we assume that each column of $\boldsymbol{w}_Y$ and $\mathbf{Z}_Y$ is centered. The least squares estimates for $\boldsymbol{\phi}_X$ and $\boldsymbol{\phi}_Y$ are

$$\widehat{\boldsymbol{\phi}}_X = \left(\mathbf{Z}_X^T\mathbf{Z}_X\right)^{-1}\mathbf{Z}_X^T\boldsymbol{w}_X$$

$$\text{and} \quad \widehat{\boldsymbol{\phi}}_Y = \left(\mathbf{Z}_Y^T\mathbf{Z}_Y\right)^{-1}\mathbf{Z}_Y^T\boldsymbol{w}_Y, \tag{A22}$$

and the corresponding sum square errors are

$$\mathrm{SSE}_X = \|\boldsymbol{w}_X - \mathbf{Z}_X\widehat{\boldsymbol{\phi}}_X\|^2$$

$$\text{and} \quad \mathrm{SSE}_Y = \|\boldsymbol{w}_Y - \mathbf{Z}_Y\widehat{\boldsymbol{\phi}}_Y\|^2, \tag{A23}$$

where $\|\boldsymbol{a}\| = \boldsymbol{a}^T\boldsymbol{a}$ denotes the Euclidean norm of vector $\boldsymbol{a}$. The sum square errors are related to the estimated variances as

$$\mathrm{SSE}_X = (N_X - p)\overline{\sigma}_X^2 = \nu_X\hat{\sigma}_X^2, \tag{A24}$$

$$\mathrm{SSE}_Y = (N_Y - p)\overline{\sigma}_Y^2 = \nu_Y\hat{\sigma}_Y^2. \tag{A25}$$

Following the standard theory of least squares estimation for the general linear model, Brockwell and Davis (1991;

Sect. 8.9) suggest that the sum square errors have the following approximate chi-squared distributions:

$$\frac{\mathrm{SSE}_X}{\sigma_X^2} \sim \chi^2_{\nu_X}, \quad \text{and} \quad \frac{\mathrm{SSE}_Y}{\sigma_Y^2} \sim \chi^2_{\nu_Y}. \tag{A26}$$

Because X_t and Y_t are independent, the associated sum square errors are independent, hence under H_0, the statistic

$$F_\sigma = \frac{\hat{\sigma}_X^2}{\hat{\sigma}_Y^2} \tag{A27}$$

has an F-distribution with (ν_X, ν_Y) degrees of freedom, as stated in Eq. (20). Furthermore, if H_0 is true, then

$$\begin{aligned} \mathrm{SSE}_\sigma &= \mathrm{SSE}_X + \mathrm{SSE}_Y = (N_X + N_Y - 2p)\overline{\sigma}^2 \\ &= (\nu_X + \nu_Y)\hat{\sigma}^2 \end{aligned} \tag{A28}$$

has a (scaled) chi-squared distribution with $\nu_X + \nu_Y$ degrees of freedom; i.e.,

$$\frac{\mathrm{SSE}_\sigma}{\sigma^2} \sim \chi^2_{\nu_X+\nu_Y}. \tag{A29}$$

Straightforward algebra shows that D_ϕ in Eq. (A16) can be written as a function of F_σ as

$$\begin{aligned} D_\sigma(F_\sigma) =& \nu_X \log(\nu_X + \nu_Y/F_\sigma) + \nu_Y \log(\nu_X F_\sigma + \nu_Y) \\ & - (\nu_X + \nu_Y)\log(\nu_X + \nu_Y). \end{aligned} \tag{A30}$$

This is a U-shaped function with a minimum value of zero at $F_\sigma = 1$ and monotonic on either side of $F_\sigma = 1$. Note that if X_t and Y_t were swapped, then the X and Y labels would be swapped and $F_\sigma \to 1/F_\sigma$, which yields the same value of D_σ. Let F_{α,ν_X,ν_Y} denote the critical value such that $F > F_{\alpha,\nu_X,\nu_Y}$ has probability α when F has an F-distribution with (ν_X, ν_Y) degrees of freedom. Then, the $\alpha 100\,\%$ significance threshold for rejecting H_0 is

$$D_{\sigma,\alpha} = D_\sigma\left(F_{\alpha/2,\nu_X,\nu_Y}\right), \tag{A31}$$

where α on the right-hand side is *divided by 2*.

The least squares estimate of the common regression parameters $\boldsymbol{\phi}$ is

$$\begin{aligned} \widehat{\boldsymbol{\phi}} &= \left(\mathbf{Z}_X^T\mathbf{Z}_X + \mathbf{Z}_Y^T\mathbf{Z}_Y\right)^{-1}\left(\mathbf{Z}_X^T\boldsymbol{w}_X + \mathbf{Z}_Y^T\boldsymbol{w}_Y\right) \\ &= \left(\mathbf{Z}_X^T\mathbf{Z}_X + \mathbf{Z}_Y^T\mathbf{Z}_Y\right)^{-1}\left(\left(\mathbf{Z}_X^T\mathbf{Z}_X\right)\widehat{\boldsymbol{\phi}}_X \right. \\ &\quad \left. + \left(\mathbf{Z}_Y^T\mathbf{Z}_Y\right)\widehat{\boldsymbol{\phi}}_Y\right), \end{aligned} \tag{A32}$$

and the corresponding sum square error is

$$\mathrm{SSE}_{\phi,\sigma} = \|\boldsymbol{w}_X - \mathbf{Z}_X\widehat{\boldsymbol{\phi}}\|^2 + \|\boldsymbol{w}_Y - \mathbf{Z}_Y\widehat{\boldsymbol{\phi}}\|^2. \tag{A33}$$

We now express this in terms of parameter estimates of the individual AR models. To do this, we invoke the standard orthogonality relation

$$\|\boldsymbol{w}_X - \mathbf{Z}_X\widehat{\boldsymbol{\phi}}\|^2 = \|\boldsymbol{w}_X - \mathbf{Z}_X\widehat{\boldsymbol{\phi}}_X\|^2 + \|\mathbf{Z}_X(\widehat{\boldsymbol{\phi}}_X - \widehat{\boldsymbol{\phi}})\|^2, \tag{A34}$$

$$\|\boldsymbol{w}_Y - \mathbf{Z}_Y\widehat{\boldsymbol{\phi}}\|^2 = \|\boldsymbol{w}_Y - \mathbf{Z}_Y\widehat{\boldsymbol{\phi}}_Y\|^2 + \|\mathbf{Z}_Y(\widehat{\boldsymbol{\phi}}_Y - \widehat{\boldsymbol{\phi}})\|^2. \tag{A35}$$

It follows that

$$\begin{aligned} \mathrm{SSE}_{\phi,\sigma} =& (\mathrm{SSE}_X + \mathrm{SSE}_Y) + \|\mathbf{Z}_X(\widehat{\boldsymbol{\phi}}_X - \widehat{\boldsymbol{\phi}})\|^2 \\ & + \|\mathbf{Z}_Y(\widehat{\boldsymbol{\phi}}_Y - \widehat{\boldsymbol{\phi}})\|^2. \end{aligned} \tag{A36}$$

The difference between least squares estimates $\widehat{\boldsymbol{\phi}}_X$ and $\widehat{\boldsymbol{\phi}}$ is

$$\begin{aligned} \widehat{\boldsymbol{\phi}}_X - \widehat{\boldsymbol{\phi}} =& \left(\mathbf{Z}_X^T\mathbf{Z}_X + \mathbf{Z}_Y^T\mathbf{Z}_Y\right)^{-1}\left(\mathbf{Z}_Y^T\mathbf{Z}_Y\right) \\ & \left(\widehat{\boldsymbol{\phi}}_X - \widehat{\boldsymbol{\phi}}_Y\right), \end{aligned} \tag{A37}$$

while that between $\widehat{\boldsymbol{\phi}}_Y$ and $\widehat{\boldsymbol{\phi}}$ is

$$\begin{aligned} \widehat{\boldsymbol{\phi}}_Y - \widehat{\boldsymbol{\phi}} =& -\left(\mathbf{Z}_X^T\mathbf{Z}_X + \mathbf{Z}_Y^T\mathbf{Z}_Y\right)^{-1}\left(\mathbf{Z}_X^T\mathbf{Z}_X\right) \\ & \left(\widehat{\boldsymbol{\phi}}_X - \widehat{\boldsymbol{\phi}}_Y\right). \end{aligned} \tag{A38}$$

Substituting these expressions into Eq. (A36) and invoking Eq. (A28) yields

$$\mathrm{SSE}_{\phi,\sigma} = \mathrm{SSE}_\sigma + \left(\widehat{\boldsymbol{\phi}}_X - \widehat{\boldsymbol{\phi}}_Y\right)^T \boldsymbol{\Sigma}_{\mathrm{HM}}\left(\widehat{\boldsymbol{\phi}}_X - \widehat{\boldsymbol{\phi}}_Y\right), \tag{A39}$$

where

$$\boldsymbol{\Sigma}_{\mathrm{HM}} = \left(\left(\mathbf{Z}_X^T\mathbf{Z}_X\right)^{-1} + \left(\mathbf{Z}_Y^T\mathbf{Z}_Y\right)^{-1}\right)^{-1}. \tag{A40}$$

$\boldsymbol{\Sigma}_{\mathrm{HM}}$ is proportional to the harmonic mean of two covariance matrices. This matrix arises naturally if we recall the fact that the sample estimates of the parameters have the following distributions:

$$\begin{aligned} & \widehat{\boldsymbol{\phi}}_X \sim \mathcal{N}\left(\boldsymbol{\phi}_X, \left(\mathbf{Z}_X^T\mathbf{Z}_X\right)^{-1}\sigma_X^2\right) \\ & \text{and} \quad \widehat{\boldsymbol{\phi}}_Y \sim \mathcal{N}\left(\boldsymbol{\phi}_Y, \left(\mathbf{Z}_Y^T\mathbf{Z}_Y\right)^{-1}\sigma_Y^2\right). \end{aligned} \tag{A41}$$

Therefore, under H_0,

$$\widehat{\boldsymbol{\phi}}_X - \widehat{\boldsymbol{\phi}}_Y \sim \mathcal{N}\left(\mathbf{0}, \left(\left(\mathbf{Z}_X^T\mathbf{Z}_X\right)^{-1} + \left(\mathbf{Z}_Y^T\mathbf{Z}_Y\right)^{-1}\right)\sigma^2\right), \tag{A42}$$

hence

$$\frac{\left(\widehat{\boldsymbol{\phi}}_X - \widehat{\boldsymbol{\phi}}_Y\right)^T \boldsymbol{\Sigma}_{\mathrm{HM}}\left(\widehat{\boldsymbol{\phi}}_X - \widehat{\boldsymbol{\phi}}_Y\right)}{\sigma^2} \sim \chi_p^2. \tag{A43}$$

Again, by analogy with the standard theory of least squares estimation for the general linear model, SSE_σ and the quadratic form in Eq. (A39) are approximately independent of each other. Thus, if H_0 is true, then

$$F_{\phi|\sigma} = \frac{1}{p}\frac{\left(\widehat{\boldsymbol{\phi}}_X - \widehat{\boldsymbol{\phi}}_Y\right)^T \boldsymbol{\Sigma}_{\mathrm{HM}}\left(\widehat{\boldsymbol{\phi}}_X - \widehat{\boldsymbol{\phi}}_Y\right)}{\hat{\sigma}^2} \tag{A44}$$

has an approximate F-distribution with $(p, \nu_X + \nu_Y)$ degrees of freedom, as indicated in Eq. (21). Using the identity

$$\mathrm{SSE}_{\phi,\sigma} = (N_X + N_Y - 2p)\overline{\overline{\sigma}}^2 = (\nu_X + \nu_Y)\hat{\hat{\sigma}}^2, \tag{A45}$$

it follows that

$$\hat{\hat{\sigma}}^2 = \hat{\sigma}^2 + \frac{\left(\widehat{\boldsymbol{\phi}}_X - \widehat{\boldsymbol{\phi}}_Y\right)^T \boldsymbol{\Sigma}_{\mathrm{HM}} \left(\widehat{\boldsymbol{\phi}}_X - \widehat{\boldsymbol{\phi}}_Y\right)}{\nu_X + \nu_Y}, \tag{A46}$$

and therefore

$$\begin{aligned} D_{\phi|\sigma}(F_{\phi|\sigma}) &= (\nu_X + \nu_Y)\log\left(\frac{\hat{\hat{\sigma}}^2}{\hat{\sigma}^2}\right) \\ &= (\nu_X + \nu_Y)\log\left(1 + \frac{pF_{\phi|\sigma}}{\nu_X + \nu_Y}\right). \end{aligned} \tag{A47}$$

It is clear that $F_{\phi|\sigma}$ is non-negative, and therefore $D_{\phi|\sigma}$ is non-negative. Furthermore, $D_{\phi|\sigma}$ is a monotonic function of $F_{\phi|\sigma}$. Therefore, the $\alpha 100\,\%$ significance threshold for rejecting H_0 is

$$D_{\phi|\sigma,\alpha} = D_{\phi|\sigma}(F_{\alpha,\phi|\sigma}). \tag{A48}$$

It remains to define the critical value for rejecting H_0 based on $D_{\phi,\sigma}$. D_σ and $D_{\phi|\sigma}$ are independent because D_σ depends on the ratio of χ_X^2 and χ_Y^2, while $D_{\phi|\sigma}$ depends on the sum $\chi_X^2 + \chi_Y^2$, and the ratio and sum are independent as a consequence of Lukacs' proportion-sum independence theorem (Lukacs, 1955). Therefore, we may sample from the F-distributions (20) and (21) and then use Monte Carlo techniques to estimate the upper α100th percentile of $D_{\phi,\sigma} = D_\sigma + D_{\phi|\sigma}$ under H_0.

Code availability. An R function for performing the test proposed in this paper is available from the authors upon request.

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. We thank Martha Buckley, Robert Lund, and two anonymous reviewers for comments that led to improvements in the presentation of this work. This research was supported primarily by the National Oceanic and Atmospheric Administration (NA16OAR4310175). The views expressed herein are those of the authors and do not necessarily reflect the views of these agencies.

Financial support. This research has been supported by the NOAA (grant no. NA16OAR4310175).

Review statement. This paper was edited by Seung-Ki Min and reviewed by two anonymous referees.

References

Box, G. E. P., Jenkins, G. M., and Reinsel, G. C.: Time Series Analysis: Forecasting and Control, Wiley-Interscience, 4th Edn., 2008.

Brockwell, P. J. and Davis, R. A.: Time Series: Theory and Methods, Springer Verlag, 2nd Edn., 1991.

Brockwell, P. J. and Davis, R. A.: Introduction to Time Series and Forecasting, Springer, 2002.

Buckley, M. W. and Marshall, J.: Observations, inferences, and mechanisms of the Atlantic Meridional Overturning Circulation: A review, Rev. Geophys., 54, 5–63, https://doi.org/10.1002/2015RG000493, 2016.

Coates, D. S. and Diggle, P. J.: Test for comparing two estimated spectral densities, J. Time Ser. Anal., 7, 7–20, 1986.

Grant, A. J. and Quinn, B. G.: Parametric Spectral Discrimination, J. Time Ser. Anal., 38, 838–864, https://doi.org/10.1111/jtsa.12238, 2017.

Izenman, A. J.: Modern Mutivariate Statistical Techniques: Regression, Classification, and Manifold Learning, Springer, corrected 2nd printing Edn., 2013.

Jenkins, G. M. and Watts, D. G.: Spectral Analysis and its Applications, Holden-Day, 1968.

Lee, Y.-S.: Some Results on the Sampling Distribution of the Multiple Correlation Coefficient, J. Roy. Stat. Soc. Ser. B, 33, 117–130, 1971.

Ljung, G. M. and Box, G. E. P.: On a measure of lack of fit in time series models, Biometrika, 65, 297–303, 1978.

Lukacs, E.: A Characterization of the Gamma Distribution, Ann. Math. Statist., 26, 319–324, https://doi.org/10.1214/aoms/1177728549, 1955.

Lund, R., Bassily, H., and Vidakovic, B.: Testing Equality of Stationary Autocovariances, J. Time Ser. Anal., 30, 332–348, 2009.

Maharaj, E. A.: Cluster of Time Series, J. Classification, 17, 297–314, https://doi.org/10.1007/s003570000023, 2000.

Piccolo, D.: A Distance Measure for Classifying ARIMA Models, J. Time Ser. Anal., 11, 153–164, https://doi.org/10.1111/j.1467-9892.1990.tb00048.x, 1990.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: A summary of the CMIP5 experimental design, available at: https://pcmdi.llnl.gov/mips/cmip5/experiment_design.html (last access: 6 October 2020), 2010.

Zhang, R., Sutton, R., Danabasoglu, G., Kwon, Y.-O., Marsh, R., Yeager, S. G., Amrhein, D. E., and Little, C. M.: A Review of the Role of the Atlantic Meridional Overturning Circulation in Atlantic Multidecadal Variability and Associated Climate Impacts, Rev. Geophys., 57, 316–375, https://doi.org/10.1029/2019RG000644, 2019.

Zwiers, F. W. and von Storch, H.: Taking Serial Correlation into Account in Tests of the Mean, J. Climate, 8, 336–351, 1995.