Changes in Internal Variability due to Anthropogenic Forcing: A New Field Significance Test

EMERSON LAJOIE AND TIMOTHY DELSOLE

Department of Atmospheric, Oceanic, and Earth Sciences, and the Center for Ocean–Land–Atmosphere Studies, George Mason University, Fairfax, Virginia

(Manuscript received 5 October 2015, in final form 23 March 2016)

ABSTRACT

Changes in internal variability of seasonal and annual mean 2-m temperature in response to anthropogenic forcing are quantified for a global domain using climate models driven by a twenty-first-century high-emissions scenario. While changes in variance have been quantified previously in a univariate sense, the field significance of such changes has remained unclear. This paper proposes a new field significance test for changes in variance that accounts for spatial and temporal relationships within the domain. The test proposed here uses an optimization technique based on discriminant analysis, yielding results that are invariant to linear transformations of the data and therefore independent of normalization procedures. Multiple significance tests are employed because spatial fields can differ in many ways in a multivariate space. All climate models investigated here predict significant changes in internal variability of temperature in response to anthropogenic forcing. The models consistently predict decreases to temperature variance in regions of seasonal sea ice formation and across the Southern Ocean by the end of the twenty-first century. While more than half the models also predict significant changes in variance over ENSO regions and the North Atlantic Ocean, the direction of this change is model dependent. Seasonal mean changes are remarkably similar to annual mean changes, but there are model-dependent exceptions. Some models predict future variability that is more than double their preindustrial control variability, raising questions about the adequacy of doubling uncertainty estimates to test robustness in detection and attribution studies.

1. Introduction

Temperature extremes can have serious impacts on society (IPCC 2012). Since the middle of the twentieth century, most land areas for which there are sufficient observational records have experienced increases in the frequency and intensity of warm extremes and decreases in the frequency and intensity of cold extremes (IPCC 2012; Collins et al. 2013). Many of these changes are consistent with the hypothesis that anthropogenic global warming acts to shift the distribution of temperature toward a warmer climate. In addition to a shift, the temperature distribution might also be changing its variance, but methods for quantifying global-scale changes in variance have been criticized.

For instance, Hansen et al. (2012) claimed that the distribution of globally aggregated summer temperatures has both shifted toward a higher mean and broadened. Subsequent studies have supported Hansen et al. (2012) with respect to a shifting mean, but disagree that changes in variance have contributed to observed summer mean hot extremes (Coumou and Robinson 2013; Rhines and Huybers 2013; Huntingford et al. 2013). One source of disagreement is the procedure for normalizing temperatures at different geographic locations before aggregating them to obtain a distribution. In particular, removing the mean temperature in one period based on the mean temperature of an earlier period, as done in Hansen et al. (2012), imparts a positive bias to the variance (Tingley 2012; Rhines and Huybers 2013; Sippel et al. 2015). Another complicating factor is that the number of surface stations in a geographic region has changed over time. In particular, a decline in station density implies fewer stations for averaging, which in turn leads to larger variance (Rhines and Huybers 2013). Finally, differences in trends

Corresponding author address: Emerson LaJoie, AOES/COLA, George Mason University, 112 Research Hall, MSN 2B3, Fairfax, VA 22030.
E-mail: elajoie@masonlive.gmu.edu

DOI: 10.1175/JCLI-D-15-0718.1

between different geographic regions also contribute to differences in variance. After accounting for issues related to normalization, trends, and data density, Rhines and Huybers (2013) find that changes to the variance of summer mean temperature cannot be detected. Consistent with this conclusion, Huntingford et al. (2013) find that if trends are removed by computing temperature anomalies relative to an 11-yr local running mean, then changes in variability of seasonal and/or annual mean temperature also cannot be detected in observations. Looking at a high-emissions scenario, Coumou and Robinson (2013) show that the changes in land area experiencing temperature exceedances are well fit by a Gaussian distribution that includes a shift in the mean of local temperature to warmer values, with no change in local variability.

Cold-season variability also has been investigated. Francis and Vavrus (2012) hypothesize that the decline of sea ice extent, due to Arctic warming, causes the jet stream to grow more wavy, resulting in more frequent cold extremes. However, Barnes (2013) and Screen and Simmonds (2013) showed that trends in planetary-scale waviness are sensitive to methodology.

Screen (2014) examined a different quantity—namely, zonal means of the local variance of daily temperature over Northern Hemispheric land—and concluded that temperature variance had decreased since 1979 for fall, winter, and spring. Screen (2014) argues that this decrease in variance is caused by Arctic amplification. Specifically, cold extremes in the Northern Hemisphere are invariably associated with winds that blow from the north. Arctic amplification, however, increases temperatures more in the Arctic than at low latitudes, thereby reducing cold advection by northerly winds. Consequently, Arctic amplification causes the coldest days to warm faster than the warmest days, thus reducing variance. Consistent with this mechanism, climate models project less variable land temperatures in northern latitudinal bands during fall, winter, and spring, and models with stronger Arctic amplification tend to exhibit stronger decreases in variance (Screen 2014). Other independent studies support this conclusion. For instance, Huntingford et al. (2013) showed that climate models predict, on average, a decrease in total variability of annual mean temperature in high-emissions scenarios relative to an 11-yr running mean, with some of this decrease associated with reductions in sea ice cover. Also, Boer (2009) showed that climate models predict, on average, that the variance of temperature anomalies (relative to a low-order polynomial in time) will decrease in midlatitudes and increase slightly in the tropics.

As is the case with the studies discussed above, temperature anomalies at different geographic locations are often combined using spatial aggregation or averaging methods in order to draw a single conclusion about an overall change in variance. In addition to the loss of information about local changes in variance, these methods must also standardize temperature anomalies to remove local differences in the means, variances, and trends prior to aggregating or averaging. Unfortunately, there is no unique normalization procedure, so criticisms can be raised about any chosen normalization procedure.

An alternate approach to combining data is to compute local changes in variance and then display maps of those changes. While this approach preserves information about local changes, it leads to a field significance problem in which the likelihood of the computed field of changes needs to be quantified relative to the null hypothesis of no local change in variance. Standard field significance techniques (e.g., Livezey and Chen 1983) are designed for correlation maps, and it is not clear how to apply them to variance ratio maps to test for field significance.

In this paper, we propose a new field significance test that quantifies the likelihood that a field of computed changes in variance could have occurred by random chance under a null hypothesis of no change in local variance or covariance within the field. This test is invariant to normalization procedures or any other affine transformation of the data. While our test avoids certain problems that arise in spatial aggregation, comprehensively accounts for temporal and spatial relationships within the domain, and has a well-defined significance measure, it unfortunately requires severely restricting the dimension of the state space. To validate the methodology, we apply our test to a selection of CMIP5 simulations and demonstrate that it gives robust results. Overall, climate models project significant changes to the internal variability of annual and seasonal mean temperature in the twenty-first century. The precise datasets we use are described in the next section, and details of our methodology are discussed in section 3. We conclude with a discussion of our results, their implications for climate change, and a discussion of other results relating to the new field significance test.

2. Data

We examine climate model simulations from phase 5 of the Coupled Model Intercomparison Project (CMIP5). Two types of simulations are analyzed: preindustrial control runs, in which the forcings do not change from year to year, and projections based on the representative concentration pathway 8.5 (RCP8.5), in which concentrations and emissions increase such that radiative forcing peaks at 8.5 $W m^{-2}$ in 2100 (Collins et al.

TABLE 1. List of climate models used in this study. Included in the table are the modeling center, the long-form model name, and the short-form model name created for this investigation and referenced herein. (Acronym expansions are available online at http://www.ametsoc.org/PubsAcronymList.)

Cimate models		
Model center	Model name	Short name
Canadian Centre for Climate Modelling and Analysis	CanESM2	CCCma
Centre National de Recherches Météorologiques–Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique	CNRM-CM5	CNRM
L'Institut Pierre-Simon Laplace	IPSL-CM5A-LR	IPSL
Atmosphere and Ocean Research Institute (The University of Tokyo), National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology	MIROC5	MIROC5
Met Office Hadley Centre	HadGEM2-ES	HadGEM2
Max Planck Institute for Meteorology	MPI-ESM-LR	MPI
National Center for Atmospheric Research	CCSM4	NCAR

2013). We analyzed seasonal and annual mean 2-m temperature fields. We evaluated only the models with both a 500-yr-long preindustrial control simulation and a three-member ensemble from the model's corresponding RCP8.5 simulations that covered the 90-yr period from 2006 to 2095. These criteria resulted in a selection of seven global climate models (see Table 1 for details). All data were interpolated onto a common $5^\circ \times 5^\circ$ grid, yielding 2592 total grid points for each model. The control simulations were detrended to remove the effects of model drift. To demonstrate robustness, we divided the simulations into two equal halves and performed some of our analyses on each half separately. For instance, the 500-yr control simulations were separated into a first and second half, each half containing 250 years of data. Similarly, each of the three 90-yr RCP8.5 members were divided into a first half (45 yr) and second half (also 45 yr). This yielded 135 total years (3×45) for each half of an RCP8.5 simulation.

3. Methodology

To quantify the response of internal variability to anthropogenic forcing, we assume that a climate variable, such as 2-m temperature $\mathbf{t}$ can be modeled as

$$\mathbf{t} = \mathbf{F}a + \mathbf{u}, \tag{1}$$

where $\mathbf{F}$ is the forced response pattern, a is the corresponding amplitude, and $\mathbf{u}$ is a random term representing internal (or unforced) variability. The statistical model (1) is commonly used in climate change detection and attribution studies, and it assumes that internal variability is independent in time, has a known distribution, and is additive relative to the variability of the forced component (e.g., Allen and Tett 1999; Jones et al. 2013; Imbers et al. 2014).

To the extent that this statistical model is correct, changes in internal variability due to anthropogenic forcing can be estimated using ensemble techniques and then compared with estimates of internal variability simulated from preindustrial control runs. To see this, consider an ensemble of simulations initialized from different states but driven by the same forcings. For such an ensemble, $\mathbf{F}a$ in (1) is the same for different ensemble members; hence, the difference that results from ensemble member minus ensemble mean yields cancellation of the forced response (i.e., cancellation of $\mathbf{F}a$), leaving an estimate of internal variability. The residual has slightly less variance than the true internal variability because the ensemble mean that is removed also contains some internal variability as a result of the finite ensemble size. In this paper, we evaluate the ensemble mean from a three-member RCP8.5 simulation and subtract it from each member. We refer to the future emissions scenario as "21C." Let $t^{21C}_{s,y,e}$ be a climate variable from the eth ensemble member at the sth spatial grid point and the yth year. Then, a (slightly damped) realization of internal variability in the twenty-first century is

$$U^{21C}_{s,y,e} = t^{21C}_{s,y,e} - [t]^{21C}_{s,y}, \tag{2}$$

where the ensemble mean of the twenty-first century simulations is

$$[t]^{21C}_{s,y} = \frac{1}{E}\sum_{e=1}^{E} t^{21C}_{s,y,e}, \tag{3}$$

and E is the total ensemble size. For the preindustrial control runs, let $t^{ctr}_{s,y}$ be a climate variable from the preindustrial control simulation (ctr) at the sth spatial grid point and the yth year. Since the forcing does not change from year to year in the preindustrial control run, $\mathbf{F}a$ in

(1) is constant. Therefore, internal variability in the control run can be estimated from the residual of the time mean. Again, the residual has slightly less variance than the true internal variability because the time mean contains internal variability as a result of finite sample size. A (slightly damped) realization of internal variability in the absence of anthropogenic forcing is

$$U^{\mathrm{ctr}}_{s,y} = t^{\mathrm{ctr}}_{s,y} - \bar{t}^{\mathrm{ctr}}_{s}, \tag{4}$$

where the climatological mean is

$$\bar{t}^{\mathrm{ctr}}_{s} = \frac{1}{Y_{\mathrm{ctr}}} \sum_{y=1}^{Y_{\mathrm{ctr}}} t^{\mathrm{ctr}}_{s,y}. \tag{5}$$

a. Univariate test for changes to internal variability

At each grid point s, we can assess if anthropogenic forcing changes variability by testing the null hypothesis that (2) and (4) were drawn from populations with equal variances. Standard analysis of variance (ANOVA) techniques show that an unbiased estimate of the variance from a realization of internal variability in a twenty-first century simulation can be determined from

$$\sigma^2_{s,21\mathrm{C}} = \frac{1}{Y_{21\mathrm{C}}(E-1)} \sum_{y=1}^{Y_{21\mathrm{C}}} \sum_{e=1}^{E} (U^{21\mathrm{C}}_{s,y,e})^2, \tag{6}$$

and an unbiased estimate of variance in the control simulations is

$$\sigma^2_{s,\mathrm{ctr}} = \frac{1}{Y_{\mathrm{ctr}}-1} \sum_{y=1}^{Y_{\mathrm{ctr}}} (U^{\mathrm{ctr}}_{s,y})^2. \tag{7}$$

If the samples are independent and identically distributed (iid) as a Gaussian (or normal distribution), then standard statistical theory states that the statistic

$$F_s = \frac{\sigma^2_{s,21\mathrm{C}}}{\sigma^2_{s,\mathrm{ctr}}} \tag{8}$$

has an F distribution with $Y_{\mathrm{ctr}} - 1$ and $Y_{21\mathrm{C}}(E-1)$ degrees of freedom. The above statistic will be called the "21C noise to control ratio." A priori we do not know in which direction the internal variability may change, so we use a two-tailed test to determine the significance of the ratio in (8). If the null hypothesis is true, this ratio should be close to one, whereas values far from one indicate that the variances differ and suggest that anthropogenic forcing changes internal variability. The statistic (8) is univariate because it compares variances at a single grid point. The spatial distribution of the variance ratios F_s can be visualized as a field of individual ratios, which we refer to as a variance ratio map [see, e.g., Fig. 4 in Boer (2009) and Fig. 1 explained in section 4]. The validity of assuming a Gaussian distribution is addressed in section 4.

b. Null hypothesis for field significance

The F test defined above determines the significance of changes in internal variability at individual grid points. We want to quantify the likelihood that a collection of variance ratios F_s could have occurred by random chance under a null hypothesis of no change in local variance or no changes in covariance between the grid points. The central issue in testing field significance is accounting for dependencies between grid points. These dependencies can be quantified by computing a covariance matrix from the data. We propose that the appropriate null hypothesis for testing differences between fields of variances is that the respective distributions have the same covariance matrix. Thus, if $\mathbf{\Sigma}_{21\mathrm{C}}$ and $\mathbf{\Sigma}_{\mathrm{CTR}}$ are the covariance matrices of internal variability in the twenty-first century and preindustrial control simulations, then our null hypothesis can be written as

$$H_0: \mathbf{\Sigma}_{21\mathrm{C}} = \mathbf{\Sigma}_{\mathrm{CTR}}. \tag{9}$$

In each covariance matrix, the diagonal elements give the variances, and the off-diagonal elements quantify the degree of dependence (covariance) between the grid points in the field. We include the off-diagonal elements in our null hypothesis because these define the dependencies across grid points that are essential to determining field significance.

Testing hypotheses about covariance matrices requires estimating the covariance matrix itself. Unfortunately, sample covariance matrices estimated from gridded data will be singular because the number of grid points far exceeds the number of samples. Singular covariance matrices present complications that we prefer to avoid. Accordingly, we project the data onto a smaller dimensional subspace of T leading empirical orthogonal functions (EOFs). We denote the EOFs by the matrix

$$\dot{\mathbf{E}} = (\mathbf{e}_1 \quad \mathbf{e}_2 \quad \cdots \quad \mathbf{e}_T), \tag{10}$$

where $\mathbf{e}_j$ denotes the jth EOF, and the overhead dot ($\dot{\cdot}$) indicates that the EOFs have been truncated at T vectors (denoted with a superscript T). Time series for the EOFs are derived from the pseudoinverse $\dot{\mathbf{E}}^i$, which has the property

$$\dot{\mathbf{E}}^T \dot{\mathbf{E}}^i = \mathbf{I}. \tag{11}$$

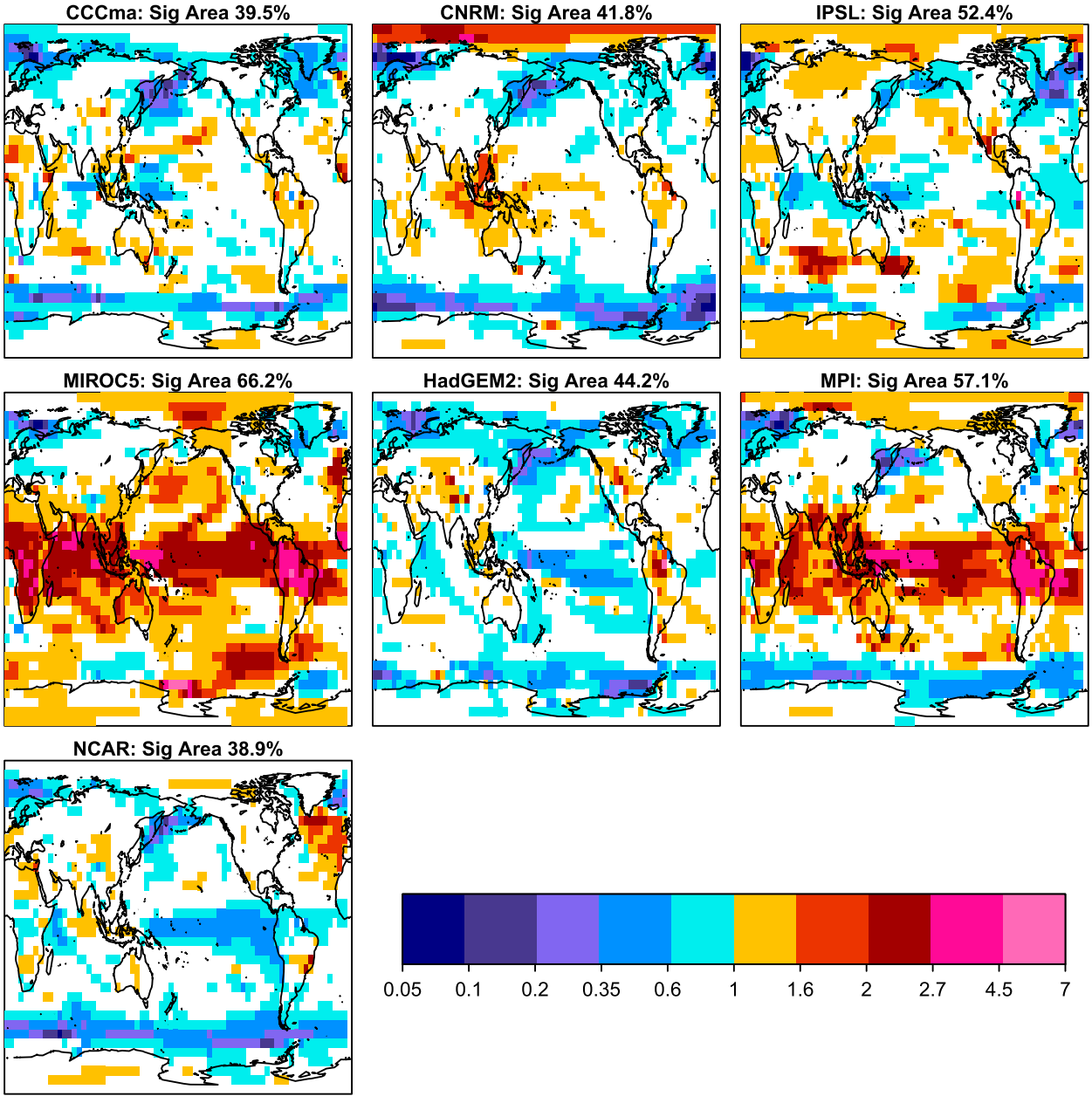

FIG. 1. Change in internal variability of annual mean 2-m temperature due to anthropogenic forcing, as quantified by the local ratio of variance in the twenty-first-century to preindustrial control internal variability in each model [(8)]. The variance of internal variability during the twenty-first century is computed from residuals about the ensemble mean of a three-member ensemble using a high-emissions scenario (RCP8.5) for the 90-yr period from 2006 to 2095. The variance of internal variability for preindustrial forcing is computed from the same model's 500-yr control simulation. A ratio >1 indicates internal variability increases in the twenty-first century. Insignificant values (according to the *F*-test distribution at a 10% significance level) are masked out (i.e., not colored). The percent area containing significant variance ratios is indicated in the title of each panel.

It is convenient to use the matrix notation

$$(\mathbf{U}^{21C})_{s,y'} = U^{21C}_{s,y,e}, \tag{12}$$

where the ensemble members have been ''stacked'' according to $y' = y + Y(e-1)$. Similarly,

$$(\mathbf{U}^{ctr})_{s,y} = U^{ctr}_{s,y}. \tag{13}$$

Then, projecting the pseudoinverse on a twenty-first century simulation yields a matrix with Y_{21C} years and T time series, referred to as principal components (PCs):

$$(\mathbf{U}^{21C})^T\dot{\mathbf{E}}^i = \mathbf{F}_{21C}. \tag{14}$$

In addition, projecting the pseudoinverse on a control simulation yields a matrix with Y_{ctr} years and T components:

$$(\mathbf{U}^{ctr})^T\dot{\mathbf{E}}^i = \mathbf{F}_{ctr}. \tag{15}$$

Using overhead tildes to denote quantities in EOF space, and recalling that estimates of a realization of internal variability have zero sample mean, we define the sample covariance matrix of internal variability in a twenty-first century simulation from the PCs in (14) as

$$\tilde{\mathbf{\Sigma}}_{21C} = \frac{1}{Y_{21C}(E-1)}\mathbf{F}_{21C}^T\mathbf{F}_{21C} \tag{16}$$

and the sample covariance matrix of a control simulation from the PCs given by (15)

$$\tilde{\mathbf{\Sigma}}_{CTR} = \frac{1}{Y_{ctr}-1}\mathbf{F}_{ctr}^T\mathbf{F}_{ctr}. \tag{17}$$

We pool the first half of the twenty-first century and preindustrial control runs to derive the EOFs. We then project those EOFs on the second halves of each dataset, yielding PCs for the second halves (see section 2 for details on how the data were divided). Dividing the datasets, and using both the twenty-first century and control runs to derive EOFs, allows us to check for robustness and account for possible biases introduced by the EOFs.

c. *Discriminant analysis*

There are many ways in which two covariance matrices can differ. We apply an optimization technique known as discriminant analysis that finds a linear combination of variables that maximizes a variance ratio. If the matrices are equal, then all linear combinations have equal variances, whereas if the covariance matrices are not equal, then discriminant analysis can diagnose the difference in an insightful manner. Let the weighting coefficients for a linear combination of variables be $\tilde{\mathbf{q}}$. Then, the ratio of variances between the twenty-first century and preindustrial contol simulations is

$$\lambda = \frac{\tilde{\mathbf{q}}^T\tilde{\mathbf{\Sigma}}_{21C}\tilde{\mathbf{q}}}{\tilde{\mathbf{q}}^T\tilde{\mathbf{\Sigma}}_{CTR}\tilde{\mathbf{q}}}. \tag{18}$$

If the null hypothesis of equal covariances is true, then $\lambda = 1$ for all possible $\tilde{\mathbf{q}}$. Conversely, if the null is *not* true, then $\lambda \neq 1$ for at least one $\tilde{\mathbf{q}}$. We seek the weighting coefficients that makes λ an extremum, which can be found by solving $\partial\lambda/\partial\tilde{\mathbf{q}} = 0$. This gives

$$\frac{\partial\lambda}{\partial\tilde{\mathbf{q}}} = \frac{2\tilde{\mathbf{\Sigma}}_{21C}\tilde{\mathbf{q}}}{\tilde{\mathbf{q}}^T\tilde{\mathbf{\Sigma}}_{CTR}\tilde{\mathbf{q}}} - 2\frac{\tilde{\mathbf{q}}^T\tilde{\mathbf{\Sigma}}_{21C}\tilde{\mathbf{q}}}{(\tilde{\mathbf{q}}^T\tilde{\mathbf{\Sigma}}_{CTR}\tilde{\mathbf{q}})^2}\tilde{\mathbf{\Sigma}}_{CTR}\tilde{\mathbf{q}}$$
$$= \frac{2}{\tilde{\mathbf{q}}^T\tilde{\mathbf{\Sigma}}_{CTR}\tilde{\mathbf{q}}}(\tilde{\mathbf{\Sigma}}_{21C}\tilde{\mathbf{q}} - \lambda\tilde{\mathbf{\Sigma}}_{CTR}\tilde{\mathbf{q}}) = 0. \tag{19}$$

Since $\tilde{\mathbf{\Sigma}}_{CTR}$ is positive definite, the derivative vanishes if

$$\tilde{\mathbf{\Sigma}}_{21C}\tilde{\mathbf{q}} = \lambda\tilde{\mathbf{\Sigma}}_{CTR}\tilde{\mathbf{q}}. \tag{20}$$

Equation (20) is a generalized eigenvalue problem. Solving this generalized eigenvalue problem for $T \times T$ covariance matrices yields T eigenvalues, or discriminant ratios, that characterize the differences between the two covariance matrices in (18). The eigenvalues can be ordered largest to smallest as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_T$. The largest eigenvalue λ_1 represents the leading discriminant ratio and gives the maximum variance ratio out of all the possible weighting vectors $\tilde{\mathbf{q}}$. Similarly, λ_T represents the trailing discriminant ratio and gives the minimum variance ratio out of all the possible weighting vectors. An important property of the eigenvalues is that they are invariant to affine transformations of the data. Centering and normalizing by a standard deviation are special cases of affine transformations.

The invariance property of the discriminant ratios yields another attractive feature: the sampling distribution of the eigenvalues are independent of the mean and covariance matrix of the population. As such, significance thresholds can be estimated by straightforward Monte Carlo methods: namely, by drawing iid random variables from a standardized univariate normal distribution. However, we also use permutation techniques to derive significance thresholds; this technique relaxes normal, iid assumptions. The consistency between the two significance thresholds will indicate the appropriateness of Gaussian and iid assumptions. This issue will be discussed in more detail in section 4.

Since each eigenvalue is individually invariant to affine transformations of the data, any function of the eigenvalues also is invariant to affine transformations. As with all statistical optimization procedures, overfitting is a concern when the number of parameters being estimated (e.g., the weights) is not a small fraction of the sample size. In section 4, we describe the steps we took to guard against overfitting by testing significance under various assumptions about the population and by demonstrating that our conclusions are robust when applied to independent data.

d. Union–intersection test

A standard test for differences in covariance matrices is the union–intersection test (Flury 1985). This test is based on testing the significance of the leading and trailing discriminant ratios λ_1 and λ_T. In essence, the hypothesis H_0 in (9) is rejected if λ_1 is too large or λ_T is too small. The significance thresholds depend on the distribution of the discriminant ratios under the null hypothesis, which can be computed using Monte Carlo methods. In our application, a significant λ_1 implies anthropogenic forcing increases internal variability, while a significant λ_T implies anthropogenic forcing decreases internal variability. The union–intersection test is quite insightful if a change in variance is detected, because the associated eigenvector can be used to derive a spatial pattern and time series that explains the difference, thereby facilitating visualization and physical interpretation of changes to internal variability.

We will show that applying the union–intersection test to model simulations leads to conclusions that are sensitive to EOF truncation and thus difficult to interpret. More precisely, for most models, the maximum (or minimum) discriminant was marginally significant for all truncations, and this was also true for the second-, third-, and higher-order discriminants (not shown). Since the union–intersection test is based only on the leading and trailing discriminant ratios, and in particular ignores intermediate discriminant ratios, it is well suited for identifying changes in variance caused by a single component of internal variability. For example, if anthropogenic forcing causes a global-scale El Niño–Southern Oscillation (ENSO) teleconnection pattern to change variance, then the union–intersection test is well poised to detect this change. The fact that the leading or trailing discriminant ratios tend to be only marginally significant, or not significant at all depending on EOF truncation, implies that changes to internal variability in a single component are weak or nonexistent. This result, however, does not imply that internal variability does not change. For instance, numerous independent components might change their variances, but the change in any individual component might be too small to satisfy statistical significance. Therefore, we seek a test that can detect ''small'' changes in variance that might be ''spread'' across many independent components.

e. Divergence

Another measure of the difference between two covariance matrices is the following:

$$D_T = \frac{1}{2}\mathrm{tr}[(\tilde{\mathbf{\Sigma}}_{21C}^{-1} + \tilde{\mathbf{\Sigma}}_{CTR}^{-1})(\tilde{\mathbf{\Sigma}}_{CTR} - \tilde{\mathbf{\Sigma}}_{21C})]. \quad (21)$$

We will call this measure divergence. An attractive property of this measure is that it is invariant to linear transformation. Also, for Gaussian distributions, this measure can be derived from the Kullback–Leilber divergence, which itself is fundamental to a wide range of applications, including information theory, finance, coding theory, and quantum entanglement (Kullback 1968; Cover and Thomas 1991; Jaeger 2007). We use (21) to measure differences in covariance matrices even for possibly non-Gaussian distributions.

An equivalent expression for divergence can be written in terms of discriminant ratios [see Kullback (1968), chapter 9, (6.7) and note the means are zero]:

$$D_T = \frac{1}{2}\sum_{i=1}^{T}\left(\lambda_i + \frac{1}{\lambda_i} - 2\right), \quad (22)$$

where λ_i is the ith eigenvalue from (18). Notice that if the covariance matrices are equal, then all the eigenvalues equal 1 and $D_T = 0$. More specifically, the function $\lambda + 1/\lambda$ is a minimum when $\lambda = 1$ and becomes large when λ is either very large or close to zero (because the function involves both the eigenvalue and its inverse). In contrast to the union–intersection test, D_T depends on the whole spectrum of eigenvalues up to the cutoff T. Thus, changes in variance that are ''spread'' across many independent components will inflate individual eigenvalues and thereby accumulate in the sum to produce a large value of D_T. Recall that the invariance property of the individual discriminant ratios means that (22) is also invariant to affine transformations of the data.

We wish to address subtleties that are associated with using a field significance test. In essence, a field significance test is designed to determine if a *collection* of values is significant; as such, these tests are not designed to determine if any *single* grid point is significant. However, if the field significance test is repeated in independent datasets and significant changes are robust, then it would be appropriate to identify local changes in variance.

4. Results

a. Changes in internal variability of annual mean temperature due to anthropogenic forcing

Local changes in internal variability of annual mean 2-m temperature between the twenty-first-century residuals (relative to the ensemble mean) and the preindustrial control run for seven climate models are shown in Fig. 1. The changes are quantified by the ratio of variances (8) between 21C noise and control time series. Insignificant values (at the 10% significance

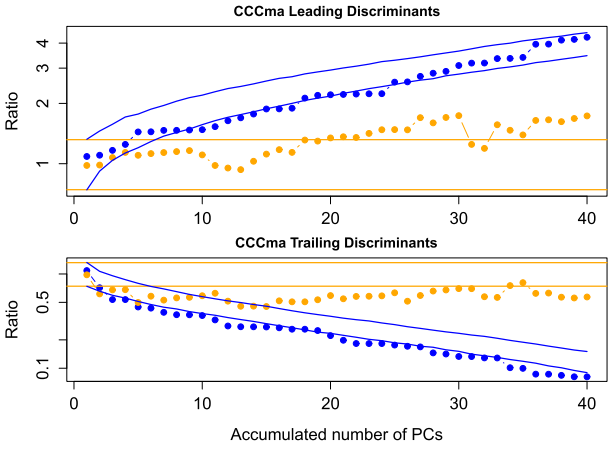

FIG. 2. The (top) maximized and (bottom) minimized variance ratios (i.e., discriminant ratios) for internal variability of annual mean 2-m temperature between twenty-first century and preindustrial control simulations from a representative model (CCCma). The results for the first half of the twenty-first century and first half of the preindustrial control run are given by the blue dotted–dashed curves, and the remaining halves are given by the orange dotted–dashed curves. Also shown are the upper and lower 5% significance thresholds computed from Monte Carlo techniques (solid blue curves) and the significance thresholds derived from an F distribution (solid orange lines).

level) are masked out. Grid points where the values are greater than 1 indicate twenty-first-century variability has increased in response to anthropogenic forcing (relative to that model's control variability at the same grid point). For instance, a value of 2 indicates that variance is projected to double relative to control variability. Conversely, values less than 1 indicate anthropogenic forcing decreases variability. In the title of each F map is the percent of total grid points deemed significant. All models predict significant change in internal variability for the RCP8.5 emissions scenario. In particular, in regions of seasonal sea ice formation (e.g., the Southern Ocean, Greenland Sea, and Bering Sea), each model consistently projects decreases in variability. This decrease is a plausible consequence of sea ice melting as a result of twenty-first-century warming: melting sea ice exposes the underlying sea surface temperature, which has less variance, owing to its larger effective heat capacity relative to sea ice. Huntingford et al. (2013) and DelSole et al. (2014) also noted variance decreases associated with areas of sea ice formation. However, the locations and directions of other changes are model dependent. For instance, most models project significant changes in variability in the tropical oceans, for example in regions of ENSO, but the direction of this change is model dependent. Similarly, the North Atlantic Ocean also exhibits significant changes in variance, but again, the direction of that change is model dependent. We find that, for the North Pole, a majority of the models indicates increases in variability. We also find several interesting, smaller-scale changes, like those in the Amazon basin; however, the scale and model-dependent direction of these changes makes them difficult to interpret.

On average, we expect to find 10% of the area of any given field of F ratios to be significant just by random chance, but clearly Fig. 1 shows many more significant changes (see percent values given in the title of each F map). While we can empirically say that the fractional area showing changes exceeds 10%, it is not clear that this is sufficient to conclude that the changes are field significant, because dependencies between the grid points have not been considered yet.

Applying the union–intersection test to the above simulations leads to results that are sensitive to the number of EOFs chosen to represent the data, thus making decisions about our null hypothesis unclear. As a representative example, we show in Fig. 2 the results of the union–intersection test for one global climate model. To check robustness, we divided our data in half and computed the test in each half separately (see section 2 for details). Results for the first halves are given by the blue dotted–dashed curve, and the second half results are given by the orange dotted–dashed curves. The top panel shows results for the leading discriminant ratio from T accumulated PCs. Significance thresholds for the first halves were derived from Monte Carlo techniques at a 10% confidence level (solid blue curves). Significance thresholds for the second halves are derived from an F distribution (solid orange curves). The bottom panel is the same as the top panel, but for the trailing discriminant ratios. The impact of overfitting can be seen in the monotonic increase (or decrease) of the significance curves as a function of the number of EOFs. The results for the leading (maximized) discriminant ratios would increase as a mathematical necessity even under a no-change hypothesis, because each additional component (EOF) provides extra freedom to fit differences in variances (i.e., overfitting). The same is true of the trailing (minimized) discriminant ratios: that is, the results would decrease as a mathematical necessity. The "marginal" results continue for EOF truncations beyond 40, but we show only up to 40 EOFs for clarity. Figure 2 shows that significance of either the maximized or minimized variance ratios is sensitive to the number of EOFs chosen. Similar results occurred for all the models considered here. These results imply that changes to internal variability in any single component are not large enough to be significant. One interpretation of this result is that large-scale components of climate variability, such as ENSO or the Pacific–North American

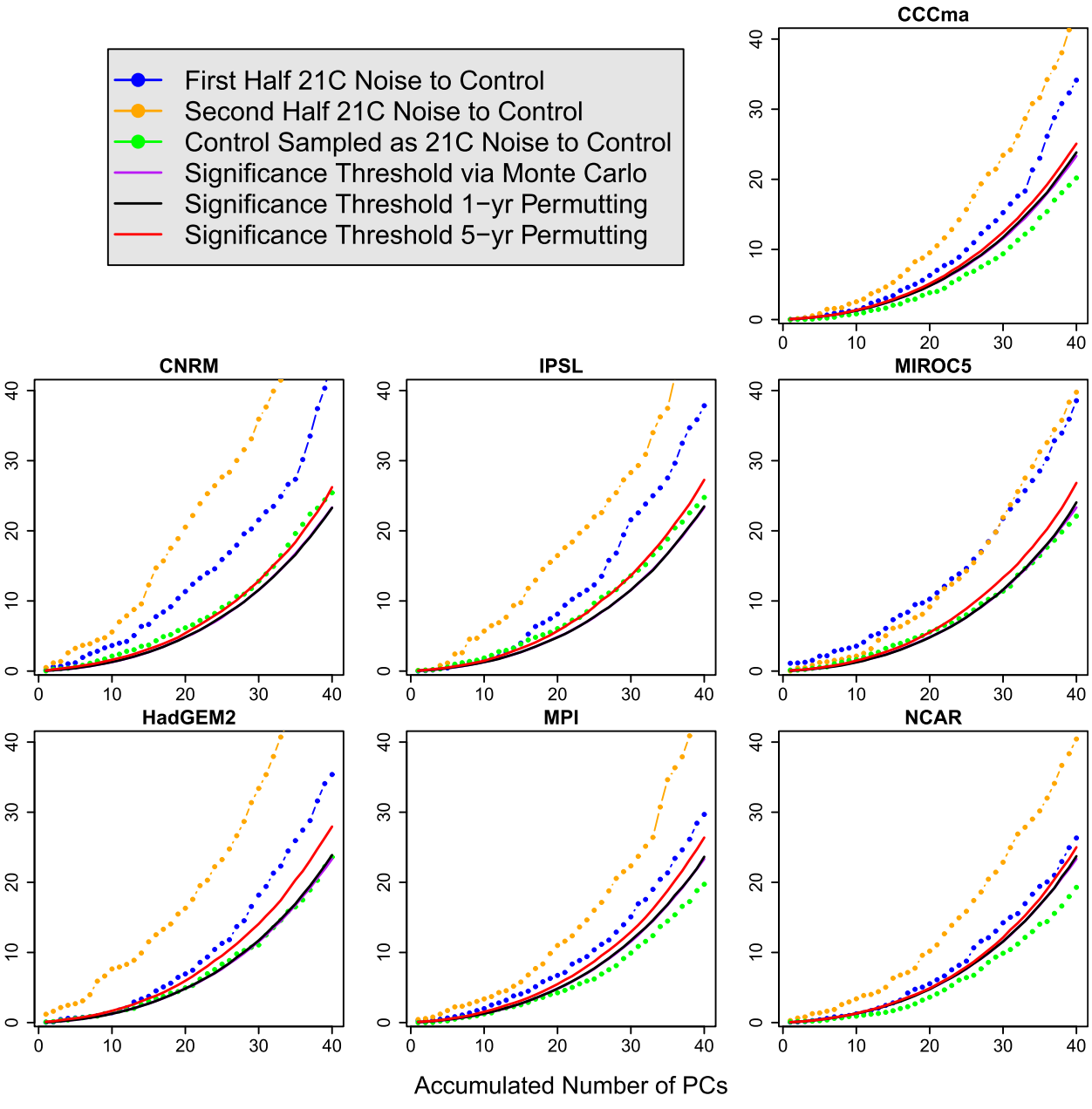

FIG. 3. The divergence D_T (y axes) of internal variability of annual mean 2-m temperature between twenty-first century and preindustrial control simulations, in each model as a function of the number of EOFs T included in the measure (x axes). The divergence for the first half of the twenty-first century and first half of the preindustrial control run are given by the blue dotted–dashed curves, and the remaining halves are given by the orange dotted–dashed curves. The green dotted–dashed curves show the divergence between different segments of a climate model's preindustrial control simulation, with dimensions that match those of the other divergences. Also shown are the upper 5% significance threshold computed from Monte Carlo techniques (purple curve) and permutation techniques: 1 yr (black curve) and 5 yr (red curve). Significant results lie above the solid curves.

(PNA) oscillation are not changing their variance in a significant way. Since the change in any individual component appears too small to satisfy statistical significance, we test if changes in variance are significant in an aggregate sense.

Accordingly, we compute the divergence of annual mean 2-m temperature between the control and RCP8.5 simulations. Again, to demonstrate robustness, we divide our data in half and compute divergence for each half separately (see section 2 for details). In Fig. 3, the

divergence results for the first halves are shown as the blue dotted–dashed curves, and the divergence results for the second halves are shown as the orange dotted–dashed curves. Both divergence curves lie above the (solid) significance curves after a sufficient number of EOFs, indicating that the F maps in Fig. 1 are field significant at a 10% significance level. This conclusion holds for EOF truncations beyond 40, but we show results only up to 40 EOFs for clarity. The impact of overfitting can be seen in the monotonic increase of the significance curves as a function of the number of EOFs. As in the union–intersection test, the divergence would increase as a mathematical necessity even under a no-change hypothesis, because each additional component (EOF) provides extra freedom to fit differences in variances (i.e., overfitting). Nevertheless, in all models the actual divergence increases much faster than that of the significance curves, indicating that the change in variance is larger than expected by random chance. Note also that the orange dotted–dashed curves in Fig. 3 are higher than the blue dotted–dashed curves for all but one model, indicating that simulated changes to internal variability are generally larger in the second half of the twenty-first century as compared to the first half (as one would expect if the changes are scaled with the degree of climate change). The green dotted–dashed curves are motivated by additional analyses that are explained in the next subsection.

The significance thresholds shown in Fig. 3 were computed three different ways to test sensitivity to assumptions about the underlying population. First, the significance thresholds were estimated by Monte Carlo methods in which independent and identically distributed random numbers were drawn from a normal distribution (purple curve). This upper 5% significance threshold is hard to see because it fits nearly perfectly beneath the significance threshold derived from a 1-yr permutation test (black curve). The permutation test randomly draws years in a control run to construct sample covariance matrices and therefore assumes only iid (i.e., it does not make a Gaussian assumption). The similarity between the purple and black significance thresholds implies that the Gaussian assumption is reasonable for our dataset. On the other hand, differences in the thresholds derived from the permutation test with 1-yr blocks (black curve) and 5-yr blocks (red curve) indicate that internal variability in the leading EOFs are autocorrelated. Choosing a 5-yr block for the permutation test was motivated by an autocorrelation study (results not shown) that revealed serial correlations ranging from 2 to 5 yr in some preindustrial control simulations. Even after accounting for autocorrelation, we find that changes to internal variability in response to anthropogenic forcing are much larger than those expected under the no-change null hypothesis even when autocorrelation is present: this is clearly indicated by the fact that the blue and orange dotted–dashed curves lie above the red curves.

A final remark: the results in Fig. 1 show that two models (MIROC5 and MPI) exhibit widespread areas in which the variance of internal variability more than doubles by the end of the twenty-first century. Some detection and attribution studies artificially inflate a model's internal variability by a factor of 2 to assess the robustness to uncertainty in the estimates of internal variability (Hegerl et al. 2007). Such studies also assume that the statistical properties of internal variability do not change in response to climate forcing. Thus, for these models, not only is the assumption of constant internal variability incorrect, but doubling the internal variability may not be sufficient to account for changes in variability due to anthropogenic forcing.

b. *Changes to internal variability of annual mean temperature in control simulations*

Previous studies find that internal variability can change on multicentennial time scales even in the absence of anthropogenic forcing (e.g., Wittenberg 2009). This hypothesis can be investigated by applying our methodology to just the preindustrial control simulations. First, at each grid point, we tested differences in variance between two, nonoverlapping 250-yr segments of a control run using the univariate F-test method described in section 3a. Recall that an F test assumes the data are Gaussian and iid. The resulting F ratios, using the annual mean 2-m temperature data, are shown in Fig. 4, and insignificant values (according to an F-test distribution at a 10% significance level) are masked out. The color scale and tick marks are the same as those in Fig. 1 and can be interpreted as the percent change in variance between the two halves of a given control run. The percent of total grid points deemed significant is given in the title of each F map. We find numerous (locally) significant changes for each model. In general, one might expect to find 10% of any given field of F ratios to be significant just by random chance. Clearly, Fig. 4 shows more changes than would be expected by random chance.

Before applying our field significance test, we wanted to determine if these differences could be explained by non-Gaussian behavior or serial correlations. With this in mind, we derived new significance thresholds for each grid point individually using a 1- and 5-yr block permutation technique. The permutation technique with 1-yr blocks only assumes iid, but not Gaussian. Permuting with a 5-yr block does not make either the

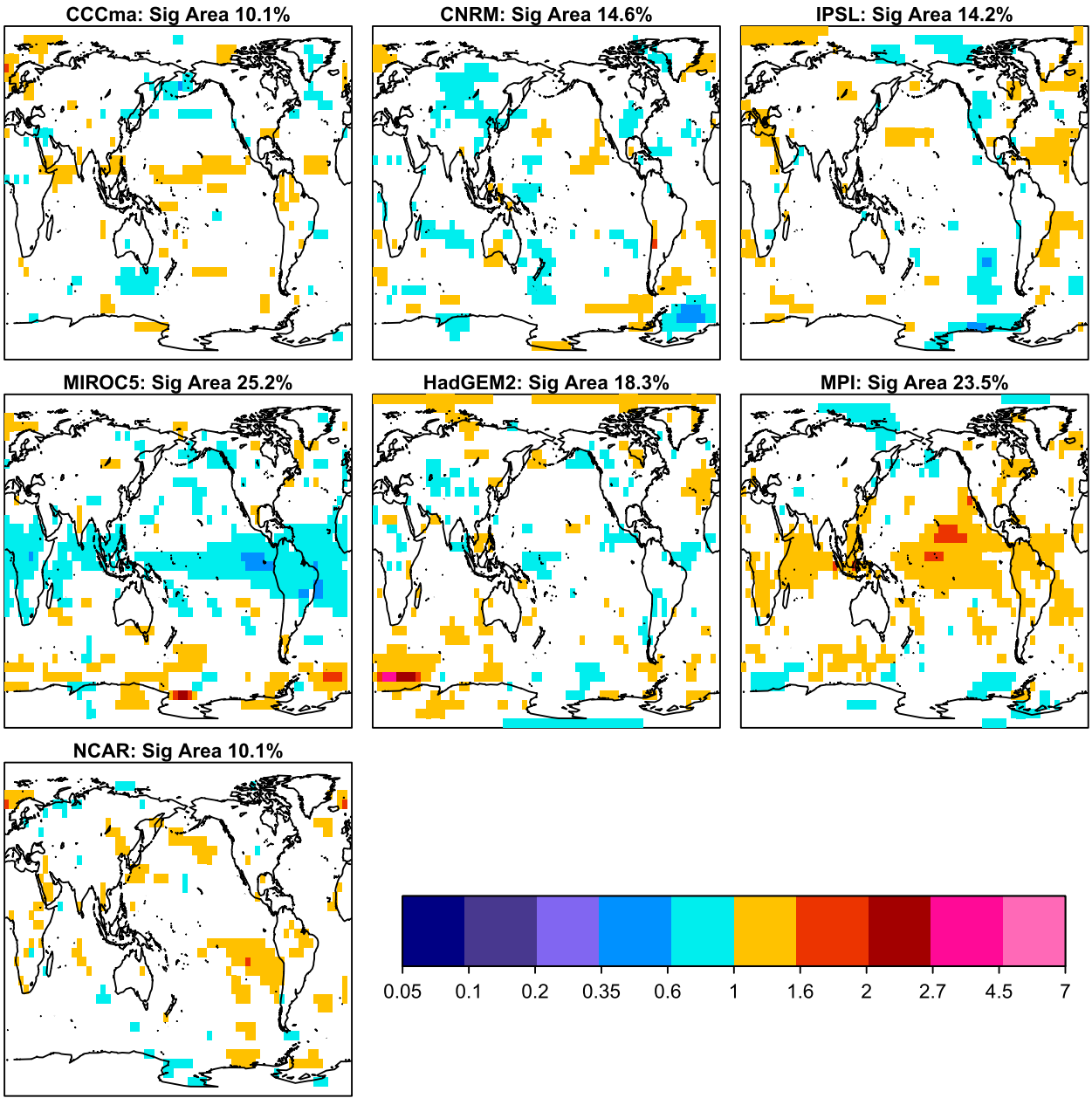

FIG. 4. Ratio of variance of internal variability of annual mean 2-m temperature between two nonoverlapping, 250-yr segments of each model's preindustrial control simulation. Insignificant values (according to an *F*-test distribution at a 10% significance level) are masked out (i.e., not colored). All but two models (MIROC5 and NCAR) have statistically significant divergence (divergence results not shown).

Gaussian or iid assumption and can account for autocorrelations. We applied the 1-yr permutation technique to the data at each grid point and found that the 1-yr permuted significance thresholds were close to those derived from an *F* distribution (not shown). Repeating permutation with a 5-yr block, we again found little difference relative to an *F* distribution. We concluded that the assumptions made by the *F* test are appropriate for these data, even after accounting for possible autocorrelations by resampling with 5-yr blocks.

We then performed our field significance test on the *F* maps in Fig. 4. We found that five of the *F* maps in Fig. 4 were field significant: namely, CCCma, CNRM, IPSL, HadGEM2, and MPI (divergence results not shown). Finding significant differences in variance within a control run helps to explain the upward sweep of the red curves in Fig. 3: that is, internal variability in the leading

EOFs derived from the preindustrial control runs *does* appear to be autocorrelated to some extent. We found that changes in the NCAR control runs were not field significant after sufficient EOF patterns were included, and the changes in MIROC5 were not field significant once a 5-yr block permutation test was computed. The remaining F maps in Fig. 4 are field significant and suggest that changes to internal variability in the absence of anthropogenic forcing exhibit an ENSO-like pattern or, more generally, large-scale changes concentrated along the equatorial Pacific Ocean with some midlatitude differences as well. However, the direction of these changes is model dependent.

Given that we have shown some models indicate internal variability changes even in the absence of anthropogenic forcing, one could question whether the changes in Fig. 1 are really caused by anthropogenic forcing or occurring naturally. The short answer is: not likely. Comparing Figs. 1 and 4, it is clear that changes in the presence of anthropogenic forcing are double or triple the changes that occur in the absence of anthropogenic forcing. However, this comparison is not perfect, because the sample sizes differ. To explore this fully, we sampled the control EOFs to mock the dimensions of and (17) and evaluated the divergence of these mock covariance matrices for each model. The resulting divergence as a function of the number of EOFs is plotted as the green dotted–dashed curve in Fig. 3. Comparing the green dotted–dashed curve with the blue or orange dotted–dashed curves clearly shows that changes in the presence of anthropogenic forcing are much greater than changes in the unforced climate system (even when the sample sizes are the same). Also note that the green dotted–dashed curve lies mostly below the red curve, indicating that the unforced changes estimated from the smaller sample size would still not be significant after accounting for autocorrelations.

c. *Changes to internal variability of seasonal mean temperature due to anthropogenic forcing*

We also looked for changes in internal variability of seasonal mean 2-m temperature [January–March (JFM), April–June (AMJ), July–September (JAS), and October–December (OND)]. The results for each season were remarkably similar with the changes in annual mean temperatures for each model, particularly for the tropics and extratropics, and for the Southern Ocean; hence, they are not shown. Of the changes in variability that were inconsistent between seasonal and annual means, most were in the Northern Hemisphere and were model specific.

We found northernmost latitudes exhibiting the greatest seasonal dependence of changes to internal variability (i.e., either increasing or decreasing seasonal mean variability). However, there were a few seasonally consistent results that were not evident from the annual mean changes. Five models agreed that decreases in variance occur about the North Pole in JAS, followed by increases in variance in OND. All the models agreed that AMJ would see decreases in variance for the North Pole region, in addition to decreases in variance in OND for the Northern Hemisphere midlatitudes (for five of the seven models). However, Northern Hemisphere JFM changes to internal variability were model dependent. Outside of the higher latitudes, there were additional changes to internal variability that were not evident in the annual mean. For example, HadGEM2 showed increased variance in Northern Hemisphere midlatitudes for JAS and JFM, especially over land. MIROC5 showed large-scale decreases in OND variability, especially over continental North America. CCCma reversed the direction of change over the Amazon basin for only OND, indicating internal variability decreases for that season. NCAR also exhibited a reversed change for the same region and season, but in the opposite direction (i.e., increases to internal variability). We found the seasonal changes to be field significant with a couple specific exceptions: for example, the first half of the JAS season was not significant for two models.

5. Summary

Current approaches to analyzing changes in climate variability often involve aggregating or averaging temperature anomalies from different geographic regions and normalizing the data to remove local differences in means, variance, and trends. There is no unique approach to normalizing data, and as such, any approach can be criticized. In addition, information about local changes is lost when the data are combined. An alternate approach is to display maps of spatially distributed changes in variance, but this presents a field significance problem; in particular, correlations in space and time need to be taken into account in order to assess the likelihood that a field of variances (and covariances) could occur under a no-change hypothesis. Capturing spatial and temporal relationships requires a multivariate approach. So, while univariate approaches do elucidate certain aspects of changes to climate variability, they have limitations.

No single test can comprehensively assess field significance, because spatial fields can differ in many ways. Certain tests detect specific departures from the null hypothesis with more power than others. In this paper, we propose a procedure for testing the field significance

of changes in variability that involves two distinct tests with complimentary approaches to detecting changes in variance. The first test, called the union–intersection test, is based on the leading or trailing discriminant ratio, which measure the maximum or minimum variance ratio out of all possible linear combinations of variables (regularized by projecting data onto a truncated set of EOFs). Because the union–intersection test is based only on the leading or trailing discriminant ratio, it is well-suited for detecting changes that occur in a single component or mode of variability. When applied to climate model simulations with and without anthropogenic forcing, the test consistently led to decisions about the null hypothesis that were sensitive to EOF truncation. Despite this, models indicated numerous local changes in variance. We interpret these results to imply that individual large-scale modes of temperature variability are not significantly changing their variance in response to increasing greenhouse gases. The second test, called the divergence test, is based on the sum of all discriminant ratios and their inverses. This test can detect small changes in variance that might be spread across many independent components. Applying this test to climate model simulations revealed significant changes in internal variability in all climate models investigated.

We applied our methodologies to investigate changes in internal variability due to anthropogenic forcing for seven climate models in the CMIP5 dataset. All the models considered here predict significant changes in internal variability of seasonal and annual mean 2-m temperature in response to anthropogenic forcing. The variance ratio maps, which characterize the local changes in variance for each model, reveal that the models consistently predict decreases in the variance of temperature in regions of seasonal sea ice formation and across the Southern Ocean in the twenty-first century (see Fig. 1). This decrease is a plausible consequence of disappearing sea ice due to global warming because melting sea ice exposes the underlying sea surface, which has a much larger effective heat capacity than sea ice owing to its coupling with the oceanic mixed layer. This interpretation also is consistent with previous research (e.g., Huntingford et al. 2013; DelSole et al. 2014; Screen 2014; Screen et al. 2014). Seasonal mean changes to internal variability are similar to annual mean changes, with noted differences primarily in the Northern Hemisphere.

More than half the models in our study also indicate significant future changes in the variance of temperature over the tropical oceans and the North Atlantic Ocean, but the sign of these changes is model dependent. ENSO's global influence on temperature and precipitation extremes makes it an important player in climate variability research. Unfortunately, our study shows that the response of ENSO to increasing greenhouse gases is highly model dependent: a result that is in accord with other studies (e.g., Vecchi and Wittenberg 2010; Collins et al. 2010).

We also find that most models exhibit significant changes in temperature variance even in the absence of anthropogenic forcing, but those changes are not as large as the changes that occur in the presence of anthropogenic forcing. In some models, the largest unforced, centennial-scale changes in variance occur along the equatorial Pacific Ocean, suggesting a connection to ENSO. This finding is consistent with previous studies that have shown significant changes to ENSO variability on centennial time scales in the absence of anthropogenic forcing (Wittenberg 2009).

We also evaluated the validity of common assumptions made about internal variability of temperature. On the one hand, consistency between permutation and Monte Carlo techniques suggests that Gaussian, iid assumptions are reasonable for seasonal and annual mean temperature data from these climate models. However, our results also suggest that the practice of doubling uncertainty estimates, as is often done in detection and attribution studies (Hegerl et al. 2007), may not be sufficient for some models in capturing the amplitude of their variability changes in response to anthropogenic forcing.

Acknowledgments. This work was sponsored by the National Science Foundation (ATM1338427), National Aeronautics and Space Administration (NNX14AM19G), the National Oceanic and Atmospheric Administration (NA09OAR4310058), and Department of Energy (DE-SC0005243). We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups for producing and making available their model output. For CMIP, the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals.

REFERENCES

Allen, M. R., and S. F. B. Tett, 1999: Checking for model consistency in optimal fingerprinting. *Climate Dyn.*, **15**, 419–434, doi:10.1007/s003820050291.

Barnes, E. A., 2013: Revisiting the evidence linking arctic amplification to extreme weather in midlatitudes. *Geophys. Res. Lett.*, **40**, 4734–4739, doi:10.1002/grl.50880.

Boer, G., 2009: Changes in interannual variability and decadal potential predictability under global warming. *J. Climate*, **22**, 3098–3109, doi:10.1175/2008JCLI2835.1.

Collins, M., and Coauthors, 2010: The impact of global warming on the tropical Pacific Ocean and El Niño. *Nat. Geosci.*, **3**, 391–397, doi:10.1038/ngeo868.

——, and Coauthors, 2013: Long-term climate change: Projections, commitments and irreversibility. *Climate Change 2013: The Physical Science Basis*, T. F. Stocker et al., Eds., Cambridge University Press, 1029–1136.

Coumou, D., and A. Robinson, 2013: Historic and future increase in the global land area affected by monthly heat extremes. *Environ. Res. Lett.*, **8**, 034018, doi:10.1088/1748-9326/8/3/034018.

Cover, T. M., and J. A. Thomas, 1991: *Elements of Information Theory*. Wiley-Interscience, 576 pp.

DelSole, T., X. Yan, P. A. Dirmeyer, M. Fennessy, and E. Altshuler, 2014: Changes in seasonal predictability due to global warming. *J. Climate*, **27**, 300–311, doi:10.1175/JCLI-D-13-00026.1.

Flury, B. N., 1985: Analysis of linear combinations with extreme ratios of variance. *J. Amer. Stat. Assoc.*, **80**, 915–922, doi:10.1080/01621459.1985.10478203.

Francis, J. A., and S. J. Vavrus, 2012: Evidence linking Arctic amplification to extreme weather in mid-latitudes. *Geophys. Res. Lett.*, **39**, L06801, doi:10.1029/2012GL051000.

Hansen, J., M. Sato, and R. Ruedy, 2012: Perception of climate change. *Proc. Natl. Acad. Sci. USA*, **109**, E2415–E2423, doi:10.1073/pnas.1205276109.

Hegerl, G. C., and Coauthors, 2007: Understanding and attributing climate change. *Climate Change 2007: The Physical Science Basis*, S. Solomon et al., Eds., Cambridge University Press, 663–745.

Huntingford, C., P. D. Jones, V. N. Livina, T. M. Lenton, and P. M. Cox, 2013: No increase in global temperature variability despite changing regional patterns. *Nature*, **500**, 327–330, doi:10.1038/nature12310.

Imbers, J., A. Lopez, C. Huntingford, and M. Allen, 2014: Sensitivity of climate change detection and attribution to the characterization of internal climate variability. *J. Climate*, **27**, 3477–3491, doi:10.1175/JCLI-D-12-00622.1.

IPCC, 2012: *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation*. C. Field et al., Eds., Cambridge University Press, 582.

Jaeger, G., 2007: *Quantum Information.* Springer, 284 pp.

Jones, G. S., P. A. Stott, and N. Christidis, 2013: Attribution of observed historical near-surface temperature variations to anthropogenic and natural causes using CMIP5 simulations. *J. Geophys. Res.*, **118**, 4001–4024, doi:10.1002/jgrd.50239.

Kullback, S., 1968: *Information Theory and Statistics.* Dover, 399 pp.

Livezey, R. E., and W. Chen, 1983: Statistical field significance and its determination by Monte Carlo techniques. *Mon. Wea. Rev.*, **111**, 46–59, doi:10.1175/1520-0493(1983)111<0046:SFSAID>2.0.CO;2.

Rhines, A., and P. Huybers, 2013: Frequent summer temperature extremes reflect changes in the mean, not the variance. *Proc. Natl. Acad. Sci. USA*, **110**, E546–E546, doi:10.1073/pnas.1218748110.

Screen, J. A., 2014: Arctic amplification decreases temperature variance in northern mid- to high-latitudes. *Nat. Climate Change*, **4**, 577–582, doi:10.1038/nclimate2268.

——, and I. Simmonds, 2013: Exploring links between Arctic amplification and mid-latitude weather. *Geophys. Res. Lett.*, **40**, 959–964, doi:10.1002/grl.50174.

——, C. Deser, and L. Sun, 2014: Reduced risk of North American cold extremes due to continued Arctic sea ice loss. *Bull. Amer. Meteor. Soc.*, **96**, 1489–1503, doi:10.1175/BAMS-D-14-00185.1.

Sippel, S., J. Zscheischler, M. Heimann, F. E. Otto, J. Peters, and M. D. Mahecha, 2015: Quantifying changes in climate variability and extremes: Pitfalls and their overcoming. *Geophys. Res. Lett.*, **42**, 9990–9998, doi:10.1002/2015GL066307.

Tingley, M. P., 2012: A Bayesian ANOVA scheme for calculating climate anomalies, with applications to the instrumental temperature record. *J. Climate*, **25**, 777–791, doi:10.1175/JCLI-D-11-00008.1.

Vecchi, G. A., and A. T. Wittenberg, 2010: El Niño and our future climate: Where do we stand? *Wiley Interdiscip. Rev.: Climate Change*, **1**, 260–270, doi:10.1002/wcc.33.

Wittenberg, A. T., 2009: Are historical records sufficient to constrain ENSO simulations? *Geophys. Res. Lett.*, **36**, L12702, doi:10.1029/2009GL038710.