A Significant Component of Unforced Multidecadal Variability in the Recent Acceleration of Global Warming

Timothy DelSole *

George Mason University, Fairfax, VA and Center for Ocean-Land-Atmosphere Studies, Calverton, MD

Michael K. Tippett

International Research Institute for Climate and Society, Palisades, NY, USA

Jagadish Shukla

George Mason University, Fairfax, VA and Center for Ocean-Land-Atmosphere Studies, Calverton, MD

* *Corresponding author address:* Timothy DelSole, 4041 Powder Mill Rd., Calverton, MD 20705.

E-mail: delsole@cola.iges.org

ABSTRACT

The problem of separating variations due to natural and anthropogenic forcing from those due to unforced internal dynamics during the twentieth century is addressed using state-of-the-art climate simulations and observations. An unforced internal component that varies on multidecadal time scales is identified by a new statistical method that maximizes integral time scale. This component, called the Internal Multidecadal Pattern (IMP), is stochastic and hence does not contribute to trends on long time scales, but can contribute significantly to short-term trends. Observational estimates indicate that the trend in the spatially averaged "well observed" sea surface temperature (SST) due to the forced component has an approximately constant value of 0.1K/decade, while the IMP can contribute about $\pm$0.08K/decade for a 30-year trend. The warming and cooling of the IMP matches that of the Atlantic Multidecadal Oscillation and is of sufficient amplitude to explain the acceleration in warming during 1977-2008 as compared to 1946-1977, in spite of the forced component increasing at the same rate during these two periods. The amplitude and time scale of the IMP are such that its contribution to the trend dominates that of the forced component on time scales less than 16 years, implying that the lack of warming trend during the past ten years is not statistically significant. Furthermore, since the IMP varies naturally on multidecadal time scales, it is potentially predictable on decadal time scales, providing a scientific rationale for decadal predictions. While the IMP can contribute significantly to trends for periods of 30 years or less, it cannot account for the 0.8°C warming trend that has been observed in the twentieth century spatially averaged SST.

1. Introduction

It is well established that the global mean surface temperature has risen by over $0.7°C$ over the last hundred years (Trenberth et al. 2007). Since global warming has been linked to rising sea levels (Bindoff et al. 2007), glacier melting (Lemke et al. 2007), Arctic sea ice retreat (Lemke et al. 2007), increasing tropical cyclone intensity (Knutson et al. 2010), and diminished snow cover (Lemke et al. 2007), the cause of this warming is of obvious concern. Many studies conclude that human activities are primarily responsible for this warming (Hegerl et al. 2007). However, the observed warming does not occur uniformly in time. For instance, the rate of global warming during 1901-2005 is about 0.075 $°C$ per decade, whereas the rate during 1981-2005 is about $0.23°C$ per decade (Trenberth et al. 2007). Also, observations indicate little to no warming during 1950-1970 and 1998-2007, despite increasing greenhouse gas concentrations (Trenberth et al. 2007).

While the reality of human induced global warming is beyond doubt, a question of intense interest is whether the recent acceleration in warming and the mid-century cooling are due to internal variability or changes in natural and anthropogenic forcing (including volcanic and solar forcing). By internal variability we mean variability that occurs in the absence of natural or anthropogenic forcing; that is, variability that occurs solely due to the internal dynamics of the coupled atmosphere-ocean-biosphere-cryosphere system. The purpose of this paper is to quantify the degree to which observed multidecadal fluctuations of spatially averaged sea surface temperature during the past century can be separated into distinct internal and forced components. While this topic has been the focus of numerous studies (Zwiers and Zhang 2003; Huntingford et al. 2006; Stone et al. 2007; Hegerl et al. 2007),

the present work explicitly identifies a significant unforced multidecadal component and separates this component from forced components using optimal spatial filtering techniques.

Detection of climate change and its attribution to external forcings requires first defining the space-time structure of the expected response of the climate system to external forcing. These forced response patterns typically are obtained from coupled atmosphere-ocean general circulation models. Because realistic climate models generate their own internal variability, identification of the forced response pattern from model simulations involves yet another signal detection problem. A typical approach is to estimate the response pattern by averaging over space, time, ensembles, or by calculating leading principal components of forced simulations. However, none of these approaches optimize detectability. In this paper, discriminant analysis is used to construct a forced response pattern that maximizes the ratio of forced variance to internal variability. A forced pattern estimated this way optimizes detection in the forced climate models and hence is likely to be detectable in observations.

Another step in detection and attribution analysis is determination of the statistical properties of internal variability. Much of the debate on global warming centers on uncertainties in the structure and magnitude of the internal variability of the real climate system. In practice, these statistical properties are estimated from climate simulations without natural or anthropogenic forcing, called control runs. After defining the forced response pattern and the statistical characteristics of internal variability, both to within unknown coefficients, generalized multivariate linear regression is then used to estimate the coefficients so as to best fit the observed record. A forced response is detected when the change in coefficient is unlikely to have occurred due to natural variability, and attributed when the change is consistent with the climate model predictions for that response, and inconsistent with the

predicted response to other plausible forcings (Hasselmann 1979, 1997; Allen and Tett 1999).

In this paper, we expand the standard detection and attribution framework by including a pattern of internal variability among the forced response patterns being investigated. This approach allows a more complete diagnosis of the role of unforced components in observed variability. The value of this approach depends on how well the internal pattern explains the variability in question. Estimation of the amplitude of internal variability proceeds in the same fashion as in standard detection analysis. Moreover, the ability to distinguish between forced and internal variability depends on the extent to which these patterns differ. However, detection and attribution are not relevant concepts for components that arise from internal variability. Instead, the concepts of skill and fidelity become relevant for internal components: skill measures the degree to which predictions of a component match observations of the component, and fidelity measures the degree to which the observed statistical properties of a component match those predicted by climate models. Methods for estimating skill and fidelity are discussed in Jolliffe and Stephenson (2003) and DelSole and Shukla (2010).

2. Identification of Internal Multidecadal Patterns

We are interested in diagnosing internal variability on decadal-to-multidecadal time scales. Unfortunately, standard statistical procedures such as principal component analysis do not decompose variables specifically by time scale. Empirical Mode Decomposition (Huang and Wu 2008) and Singular Spectrum Analysis (Ghil et al. 2002) ignore spatial correlations and hence are not optimal. Multi-channel singular spectrum analysis and extended empirical orthogonal functions (Ghil et al. 2002) often are used to decompose time series

by time scale, but are not specifically optimized for this purpose. Here we employ a novel statistical method that decomposes variables by time scale, where time scale is measured by Average Predictability Time (APT). APT is a measure of predictability that can be interpreted as a multivariate generalization of the integral time scale

$$T_2 = 1 + 2\sum_{\tau=1}^{\infty} \rho_\tau^2, \tag{1}$$

where ρ_τ is the autocorrelation function of the process and τ is time lag (DelSole and Tippett 2009a). DelSole and Tippett (2009b) show that any multivariate time series can be decomposed into an uncorrelated set of components ordered such that the first maximizes APT, the second maximizes APT subject to being uncorrelated with the first, and so on.

Unfortunately, multidecadal variability tends to be model dependent (Latif et al. 2006). To reduce this model dependence, we adopt a multi-model approach in which APT is optimized over multiple models. A potential problem with this approach is that the leading component may arise from a single dominant model or subset of models. To confirm that the leading APT component genuinely reflects a property of the entire multimodel ensemble, we verify that the component is predictable on decadal time scales in each model separately.

We examine control runs that were assessed in the Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report. Only runs that are at least 300 years long and have consistent variances are considered (see appendix sec. b for details). Using only control runs ensures that the obtained patterns are due to internal variability and not to natural or anthropogenic forcing. The component with maximum APT in these control runs is shown in the top panel of fig. 1. To facilitate comparison with observations, APT is optimized only using "well-observed" ocean model points, where "well-observed" depends on the sampling

characteristics of an observational data set (see appendix sec. a for details). The pattern is predominantly of single sign and concentrated in the North Atlantic and North Pacific. The centered time series in three representative control runs, shown in the bottom panel of fig. 1, confirm that the component fluctuates significantly on multidecadal time scales. The APT for this component is 5.2 years. Note that APT quantifies predictability time scale, not oscillatory time scale. Loosely speaking, oscillatory time scale depends on the *location* of a spectral peak, while APT depends on the *width* of a spectral peak (DelSole and Tippett 2009a). Statistical significance of APT is assessed relative to the null hypothesis that the time series is white noise when sampled every *two* years. The motivation for this hypothesis and the method for estimating the corresponding sampling distribution are discussed in the appendix sec. c. The APT of this component is found to be statistically significant in each model individually. We refer to this component as the Internal Multidecadal Pattern (IMP).

The space-time structure of the IMP is suggestive of the multidecadal variability identified in previous studies (Bjerknes 1964; Schlesinger and Ramankutty 1994; Kushnir 1994; Delworth and Mann 2000). Consistent with some of these studies, the IMP identified here has no significant correlation with concurrent atmospheric surface winds, surface pressure, or precipitation at any grid cell. The lack of correlation between major atmospheric coupling variables, and the concentration of amplitude in regions associated with Deep Water Formation, suggest that the multidecadal variations in the IMP arise from internal ocean dynamics, either as a self-sustained phenomenon or driven stochastically by the atmosphere.

Although the IMP has amplitude in two oceanic basins, the optimization procedure does not distinguish between cause and response, hence we cannot exclude the possibility that decadal variability arises in one basin and that the signal in other basins emerges as a re-

sponse. A related concern is that optimizing multimodel APT may result in a component that is a mixture of distinct phenomena in different models. For instance, one could imagine the Atlantic and Pacific structure of the IMP as resulting from some of the models having predictability in the Atlantic and others in the Pacific, but none having related predictability in both. Finally, the signal in both basins might be a statistical artifact caused by representing variability with an incomplete set of EOFs. To gain insight into these questions, we optimized APT only in the Atlantic basin, then calculated regression coefficients between the component and the SST at individual grid points. We found that optimizing APT only in the Atlantic yielded nearly the same structure in the Atlantic as seen in fig. 1. Moreover, outside the Atlantic, the regression map captured a similar positive relation in the North Pacific. Large scale positive regression coefficients can be found in the Pacific, to varying degrees, in about half the models individually. We also optimized APT only in the Pacific and obtained a pattern of the same sign globally, with positive regression coefficients in the Atlantic in about half the models individually. These calculations show that the Atlantic and Pacific signals are genuine co-varying structures and not statistical artifacts of the technique.

3. Identification of the Forced Response

To specify the response to climate forcing, we seek a pattern that describes the change in spatial structure due to natural and anthropogenic forcing, but also filters out as much internal variability as possible. To do this, we determine the pattern that maximizes the ratio of variance in the forced run to variance in the control run (details given in appendix sec. d). The mathematical technique for doing this is called discriminant analysis and has been

used previously to quantify seasonal predictability (Straus et al. 2003), decadal predictability (Venzke et al. 1999), and climate change (Ting et al. 2009; Schneider and Held 2001). Our application differs from Ting et al. (2009) in that we use the method to discriminate between forced and control runs, with no ensemble averaging, whereas Ting et al. (2009) use only forced runs to discriminate between ensemble means and deviations therefrom. To the extent that the forced response is additive and independent of internal variability, the two methods should give identical results. In practice, the two methods may give different results. Our approach takes advantage of the much larger sample sizes afforded by the control runs.

In contrast to many attribution studies, no time-lag information is used to describe the response to external forcing– the discriminant analysis is based only on spatial structure.

The pattern that maximizes the ratio of variances between forced and control runs is shown in fig. 2. Variance in each control run is measured with respect to the 300-year mean of the control run, while variance in each forced run is measured with respect to the 1901-1950 mean of the forced run. The discriminant pattern in fig. 2 is similar to that obtained in Ting et al. (2009), indicating that sampling errors are not significant. We call this pattern the forced-to-unforced discriminant. In contrast to the IMP (fig. 1), the discriminant pattern has positive anomalies in the tropics and weak or negative anomalies in the extratropics. These differences provide the basis for separating forced and internal variability.

Discriminant analysis produces an ordered set of patterns such that the first maximizes the ratio of forced-to-unforced variance, the second maximizes this ratio subject to being uncorrelated with the first, and so on. The question arises as to whether some secondary patterns should be included to capture more fully the response to climate forcing. Fig. 3 shows the variance ratios for all discriminant patterns and reveals that only the first is

clearly separated from the others. The figure also shows the 95th percentile of variance ratios determined from 1000 bootstrap samples of the control simulations, using a block length of 80 years (a large block length was chosen to capture autocorrelations and trends). The fact that only the first ratio exceeds the 95th percentile implies that the other ratios are consistent with the hypothesis of no forced response. Presumably, only one forced pattern emerges because the response to different climate forcings (e.g., fossil fuel burning, volcanic eruptions, and solar variability) tend to project on similar surface structures. Consistent with this, most attribution studies distinguish different forcings by including temporal information, vertical structure, or seasonality in the response pattern (Hegerl et al. 2007). In contrast, this study uses only horizontal spatial structure to discriminate between forced and unforced variability. Treatment of the more general detection and attribution problem, in which the response to different forcings are separated, probably would require space-time filters.

4. Results of Fingerprinting

We now fit the well-observed annual average SST at each year to a linear combination of the forced-to-unforced discriminant and the IMP. Amplitudes are chosen to approximate observations as closely as possible, where "close" is defined by a generalized distance measure that accounts for correlations in space. This procedure is equivalent to fingerprinting (Hasselmann 1979, 1997; Allen and Tett 1999; Hegerl et al. 2007) and discussed in sec. e. In contrast to previous studies that have used fingerprinting to distinguish between different forcings (e.g., anthropogenic and natural), fingerprinting is used here to distinguish between forced and internal variability, and this discrimination is based solely on spatial structure

(i.e., no temporal information is included in the response pattern). The value of including an unforced component among the forced components lies in the fact that the IMP is (1) the most predictable structure on decadal time scales in state-of-the-art climate models, and (2) of single sign and hence projects strongly on the global average. For brevity, we refer to the amplitude of the IMP projected onto the well-observed SST as the "observed IMP."

The amplitude of the forced component, expressed as a 95% confidence interval, is indicated by shading in the top panel of fig. 4. The confidence interval accounts for uncertainty due to finite sample size and missing observations (see appendix sec. f for details). The amplitude of the forced component is dominated by a secular trend, but also decreases briefly after certain major volcanic eruptions. These decreases are consistent with the fact that explosive volcanic eruptions increase sulphate aerosols in the stratosphere which in turn lead to temporary global cooling (Forster et al. 2007). These evolutionary features support the claim that this component captures the response to both anthropogenic and natural forcing.

The expected amplitude of the forced pattern is estimated by averaging the amplitude of the forced-to-unforced discriminant across the forced runs. The resulting amplitude, shown as the blue line in the upper panel of fig. 4, is smoother than the observed counterpart because fluctuations arising from internal variability are filtered out by averaging over many forced runs. The question arises as to whether the observed and predicted amplitudes agree with each other, as required for a formal attribution analysis. On a year-by-year basis, an attribution analysis is equivalent to checking that the predicted amplitude occurs within the 95% confidence interval of the observed amplitude– that is, checking that the blue curve in fig. 4 lies within the shaded region. This condition is satisfied for most years (by definition, observations are expected to fall outside the shaded region approximately 5% of the time).

In contrast, detection– that is, determining that the observed amplitude is significantly different from zero– is equivalent to checking that the 95% confidence interval for the observed amplitude does not contain zero– that is, checking that the shaded region does not intersect the zero-line. This condition is satisfied for every year after 1968. Thus, we conclude that the warming that has been observed since the 1970s is very unlikely to have occurred due to internal variability and is consistent with the warming due to anthropogenic and natural forcing predicted by models, consistent with the major conclusions of the IPCC (Hegerl et al. 2007). Note that our methodology has allowed us to draw this conclusion without taking multi-year averages, as typically performed in detection and attribution analyses (e.g., Huntingford et al. (2006); Zwiers and Zhang (2003); Stone et al. (2007)).

The amplitude of the IMP, expressed as a 68% confidence interval, is indicated by shading in the bottom panel of fig. 4. A smaller confidence interval is used because we are quantifying uncertainty rather than performing detection and attribution. A striking aspect of this time series is the multidecadal oscillations in the last 100 years. The negative anomalies during 1900-1920 and 1970-1990, and the positive anomalies during 1930-1960, closely follow the multidecadal variability identified in the Atlantic ocean by previous studies (Schlesinger and Ramankutty 1994; Kushnir 1994). To further reinforce this point, we show the annual average Atlantic Multidecadal Oscillation (AMO) index as the red curve in the bottom panel of fig. 4, which is clearly correlated with the observed IMP. These results suggest that the AMO represents variability that is dominated by internal dynamics, as suggested in previous studies (Knight et al. 2005; Zhang et al. 2007; Ting et al. 2009).

Since neither the control nor the forced runs used initial states based on observations, the random nature of internal variability means that we do not expect an internal component

in the model to predict the time evolution of the component in observations– that is, we do not expect predictions of an internal component by these models to have skill. Consequently, comparison between model and observations is confined to verifying fidelity– that is, to verifying that the statistical characteristics of the component are consistent between observations and models. Two measures of fidelity are time scale and variance. To quantify time scale, we use a sample estimate of the integral time scale (1) (see appendix sec. c for further details). For the observed IMP time series, we find $T_2 = 5.7$ years. By comparison, the mean and standard deviation of T_2 across all forced runs is 6.6 and 3.7 years, respectively. Thus, the time scale of the observed IMP is within the range of time scales predicted by models. Similarly, the variance of the observed IMP is 1.65, while the mean and standard deviation of the IMP variance across all forced runs is 1 and 0.75, respectively. Thus, the variance of the observed IMP is consistent with the range of variances predicted by models.

The squared autocorrelation function of the IMP in each control run is shown in fig. 5. Also shown is the 5% significance level of the autocorrelation based on a sample size of 300. The significance level does not account for the selection bias due to choosing the component that maximizes APT, but this bias is small owing to the large sample size of the multimodel data set (over 4200 samples), and as confirmed by splitting the data in half verified by comparing the autocorrelations to those calculated from independent control runs. The e-folding times have a mean of 7.7 years and a standard deviation of 3.5 years, but several models have significant autocorrelations after 10 years, implying that the IMP in these models can be predicted by at least a linear model on decadal time scales. These results provide a scientific rationale for decadal prediction.

5. Forcing of Internal Multidecadal Variability

Two major questions arise at this point: (1) does fingerprinting truly separate forced and unforced variability in observations, and (2) does natural and anthropogenic forcing of the twentieth century influence the evolution of the IMP? Both questions can be answered by checking for the existence of a common signal in the IMP time series of the forced runs. To detect this signal, we test whether the ensemble mean IMP varies in time. For consistency, the IMP in the forced runs is calculated the same way as the observed IMP, including using the same forced-to-unforced discriminant. The mean and 95% confidence intervals for the IMP are shown in fig. 6 (see sec. g for calculation details). About one-fifth of the confidence intervals fail to bracket zero, whereas only one-twentieth should fail to bracket zero if there were no signal. Thus, we cannot reject the hypothesis that the forcing does not influence the IMP. Nevertheless, there are three points to recognize about this conclusion. First, the ensemble mean IMP has a standard deviation of 0.24, whereas the observed IMP has a standard deviation of 1.65. Thus, to the extent forcing influences the IMP, this influence explains only about one-seventh of the observed variability in IMP. Second, the ensemble mean IMP generally decreases between 1980-2000, whereas the observed IMP generally increases during this period. Also, the ensemble mean IMP has virtually no trend between 1950-1970, whereas the observed IMP generally decreases during this period. Thus, even if the forcing influenced the IMP, this influence cannot account for the general increase in the observed IMP during 1980-2000 and the general decrease between 1950-1970. Finally, the ensemble mean IMP is autocorrelated, hence it is not appropriate to apply the significance test for the mean to each year independently. An effective time scale estimated by fitting the *ensemble*

mean IMP to a first order autoregressive model is 7.7 years (see appendix sec. g for details of this estimate). To the extent that this time scale reduces the degrees of freedom by a factor of 7.7, the error bars in fig. 6 would more than double, in which case the ensemble mean IMP in the forced runs would be indistinguishable from zero.

6. Analysis of Spatially Averaged SST

We now investigate the role of forced and internal variability in the spatially averaged SST. To do this, we consider three annual average SST data sets: (1) the observed SST, (2) the SST as reconstructed from the IMP and the forced component, and (3) the SST as reconstructed from the forced component only. The latter two data sets are sampled consistent with the missing data distribution of the first. The area-weighted spatial average of each SST data set is shown as colored dots in fig. 7. First note an apparent discontinuity at 1945. This discontinuity has been noted previously and compellingly attributed to uncorrected instrumental biases during the early 1940s (Barnett 1984; Thompson et al. 2008). Consequently, we exclude the early 1940s data from our analysis. Dividing the post-1945 era into two equal periods yields the two 32-year periods 1946-1977 and 1977-2008. The trend during these 32-year periods for the three data sets are shown as solid lines in fig. 7 and tabulated in table 1; the trend for the whole 63-year period 1946-2008 also is shown, but offset by 0.4K for clarity; 95% confidence intervals also are include in the table.

Second, the trend for the observed SST, and for the reconstruction based on IMP plus forced component, are statistically indistinguishable (i.e., their confidence intervals overlap). This agreement demonstrates that the two components capture the dominant multidecadal

fluctuations in the observed spatially average SST. Third, the trend for the observed SST is larger in the 1977-2008 period than in the 1946-1977, indicating "accelerated warming." The change in trend is significant under the assumption that the variability that remains after removing the trend is not autocorrelated. In reality, the residual variability is autocorrelated due to the IMP and nonlinear temporal forcing, raising the possibility that the two trends might not be distinguishable when autocorrelations in the residuals are taken into account. In fact, this difference can be explained by the IMP. Specifically, the trend due to only the forced component (i.e., the red line) is statistically the same in the two 32-year periods and in the 63-year period. That is, the forced part is not accelerating. Taken together, these results imply that the observed trend differs between the periods 1946-1977 and 1977-2008 not because the forced response accelerated, but because internal variability lead to relative cooling in the earlier period and relative warming in the later period.

The contribution to trends due to the IMP can be understood from a more general framework. Since the IMP is stochastic, it can contribute only random trends. The distribution of these trends can be derived analytically from the statistics of the stochastic process (see appendix sec. h for details). If the IMP is fitted to a first order autoregressive model based on its 1-year lag autocorrelation of 0.806, then the 95% confidence interval for the IMP varies with trend period length as shown in fig. 8. Note that the confidence interval increases rapidly with decreasing trend period length. For reference, the confidence interval is ± 0.169 K/decade for 16-year trends, $\pm$ 0.0776 K/decade for 32-year trends, and $\pm$ 0.031 for 64-year trends. By comparison, the trend for the forced pattern is about 0.1 K/decade, which is close to the confidence interval for the IMP trend for 25-year periods. On 10-year time scales, variability in trend due to the IMP is relatively large (e.g., ± 0.265 K/decade)

and can easily overwhelm the trend due to the forced component, although variability due to interannual variability, such as El Niño, also becomes important on this time scale. This framework provides a consistent and plausible explanation for why trends vary so strongly on 10-year time scales, and indicate that the lack of warming trend during the past 11 years (1998-2008) is not sufficient to conclude that the long term rate of warming has changed.

Consistent with the above results, all three data sets shown in fig. 7 have statistically indistinguishable trends for the 63-year period 1946-2008, indicating that internal variability can be filtered out by fitting trends over 60 or more years. Thus, in addition to optimizing forced-to-unforced variance, the forced response can be estimated from the

- Pattern of linear trends in the observations between 1850-2005.
- Pattern of linear trends in the forced runs between 1850-2000.
- Leading EOF of the ensemble mean forced runs between 1850-2000.
- Leading signal-to-noise discriminant of the forced runs between 1850-2000.

The last pattern, proposed by Ting et al. (2009), maximizes the ratio of the variance of ensemble means to the variance of the deviations therefrom. All four patterns are shown in fig. 9. The patterns generally agree on a cooling pattern in the North Atlantic, warming in the tropics, and little to no warming in the central North Pacific. The result of fitting the observed SST to each of these patterns in turn, simultaneously including the IMP, is shown in fig. 10. We see that the different time series differ only in minor ways on short time scales– i.e., the time series and confidence intervals are not sensitive to the method by which the forced pattern has been estimated. This result shows that using discriminant analysis to identify forced patterns does not enhance detectability, contrary to expectation. However,

the use of spatial optimization techniques may become more critical for smaller geographic domains or for smaller sample sizes. Note also that the trend pattern from observations does not involve a dynamical model, and hence is not affected by uncertainties in climate forcing.

It should be emphasized that while the IMP can contribute significantly to trends on time scales of around 30-years or less, it cannot account for the century-long 0.8°C warming trend observed in the spatially averaged sea surface temperature.

Acknowledgments.

We thank anonymous reviewers for thoughtful and extensive comments which lead to significant improvements in the methodology and presentation of this work. This research was supported by the National Science Foundation (ATM0332910, ATM0830062, ATM0830068), National Aeronautics and Space Administration (NNG04GG46G, NNX09AN50G), the National Oceanic and Atmospheric Administration (NA04OAR4310034, NA09OAR4310058, NA05OAR4311004), and the U. S. Department of Energy. The views expressed herein are those of the authors and do not necessarily reflect the views of these agencies.

APPENDIX

a. Data Sets

The observational data set used in this study is the HadSST2 data set compiled by the Met Office Hadley Centre and available online at http://badc.nerc.ac.uk/data/hadsst2/. This data set is an estimate of sea surface temperature (SST) from 1850 to the present, averaged on a monthly basis on a 5 degree by 5 degree grid. Grid cells with insufficient data are assigned "missing values." See Rayner et al. (2006) for further details of this data set.

The model simulations analyzed in this study are the WCRP CMIP3 multi-model dataset. These simulations were assessed in the Fourth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC). The two scenarios analyzed are: (1) radiative forcing characteristic of pre-industrial times, called "control runs" or "unforced runs" (officially designated as "PICNTRL"), and (2) anthropogenic and natural forcing characteristic of the twentieth century, called "forced runs" (officially designated as "20c3m"). The specific models used in this study are listed in table 2. Some models did not include solar and volcanic forcing in the forced runs, as indicated in table 2, so a multimodel average of the forced runs probably underestimates the response to natural forcing. The consistency of the final results with observations suggests that this underestimation is not serious.

The output of each simulation was interpolated onto the 5×5 degree grid of the HadSST2 data set to facilitate comparisons on a common grid. The time series at each ocean grid cell in each data set was averaged from January to December of each year to obtain 12-month averages. For the HadSST2 data set, only years having 10 or more months of non-missing values were averaged, otherwise the grid cell was assigned a missing value for that year.

To facilitate comparison with observations, we include a grid cell in the analysis only if 85% of the years between 1950-2005 were available in the HadSST2 data set, otherwise the grid cell was omitted from analysis. This "masking" procedure yielded a map containing 688 ocean points. All model simulations were masked in this way for consistency.

The Atlantic Multidecadal Oscillation (AMO) index used in fig. 4 was downloaded from http://www.cdc.noaa.gov/data/timeseries/AMO/. The AMO is defined as the detrended, area weighted average SST over the North Atlantic from 0 to 70°N. The index was annually averaged and scaled to fit in the figure.

b. Multimodel EOFs

Statistical optimization methods produce biased estimates due to overfitting when the number of estimated parameters is not a small fraction of the total sample size. To reduce the number of estimated parameters, we project variables onto the leading principal components (PCs) of the ensemble of control runs. Accordingly, we collected all control runs with at least 300 years of simulation, which totaled to 17 models. To avoid biasing the results toward models with longer runs or multiple ensemble members, we use only the last 300 years of the single longest control run of each climate model simulation. The 300-year mean of each simulation is then subtracted from each grid point of the corresponding control run to produce anomalies of annual mean fields. For reasons explained below, the IAP, GISS-EH, and GISS-ER models were removed from analysis. This procedure yielded 300-year anomaly time series from 14 separate climate models, giving a total of 4200 years of simulation. The specific models used in our analysis are the first 14 models listed in table 2.

The complete set of control runs obtained from the above procedure yields a 688×4200 data matrix. The principal components were computed by multiplying the anomaly at each grid point by the square root of the cosine of latitude, computing the SVD of the resulting data matrix, and then inverting the cosine weighting in the spatial fields. The cosine weighting ensures that the SVD maximizes the area weighted variance.

After computing the multi-model PC, the variance of a given PC was computed separately for each model. A standard chi-square test with 299 (=300-1) degrees of freedom indicates that the 95% confidence interval for each variance is about 18% of the variance. This interval underestimates the true interval since the time series are autocorrelated, but nevertheless provides a useful benchmark. When all 17 control simulations of length 300-years are included, the IAP model was found to have several times more variance than any other model in each of the first five PCs, suggesting that this model has significantly different space-time variability than other models and hence should not be pooled with the others. Accordingly, the IAP model was dropped from the analysis and the PCs recomputed. The newly computed PCs revealed that the GISS-EH model had more than twice as much variance as any other model, and that the GISS-ER model had less than half as much variance as any other model. Eliminating these two models and recomputing the PCs revealed that the CNRM model had 40% more variance than any other model, but otherwise the remaining control runs revealed no clear outlier models in terms of variance (i.e., the 95% confidence interval for each model overlapped with at least one other model). Given the underestimation of the confidence interval, and the multiple comparisons involved, we decided not to exclude the CNRM model from our final set of models. This procedure eliminated 3 out of 17 models, giving a total multi-model ensemble of 14 models, each of length 300 years.

It turns out that the IAP, GISS-EH, and GISS-ER models also have significant trends in the control runs. APT analysis is sensitive to such trends and including these models in the analysis produced results that were dominated by these models. Thus, these models are "outliers" not only in terms of variance, but also in their multidecadal variability.

c. Internal Multidecadal Patterns

We employ a novel procedure for identifying characteristic patterns of internal multidecadal variability. Complete details can be found in DelSole and Tippett (2009a,b). Briefly, the method is to optimize Average Predictability Time (APT), which is defined as the integral over lead time of the "signal-to-total" ratio of a forecast model, where "signal" is the variance of the ensemble mean at fixed lead time, and "total" is the corresponding total variance of the forecast ensembles. For a multivariate, stationary, Gaussian, Markov process, maximizing APT leads to the generalized eigenvalue problem

$$\left(2 \sum_{\tau=1}^{\infty} \mathbf{\Sigma}_\tau \mathbf{\Sigma}_0^{-1} \mathbf{\Sigma}_\tau^T \right) \mathbf{q} = \lambda \mathbf{\Sigma}_0 \mathbf{q}, \tag{A1}$$

where $\mathbf{q}$ is the desired projection vector, $\mathbf{\Sigma}_\tau$ is the time-lagged covariance matrix of the process, τ is the time lag, and superscript T denotes the transpose matrix. The eigenvectors provide the basis for decomposing the multivariate time series into a complete, uncorrelated set of components ordered such that the first maximizes APT, the second maximizes APT subject to being uncorrelated with the first, and so on.

In practice, the data is first projected onto the leading principal components of the control runs. (The inverse projection is somewhat subtle and discussed in Schneider and Held (2001) and DelSole and Tippett (2007).) The results are virtually independent of the number of

PCs in the range 10-100 PCs, presumably because the time series are relatively long. We choose 40 EOFs for displaying results, which is less than 1% the number of samples.

The time-lagged covariances become less certain as the time lag increases, since the amount of data available for averaging decreases with time lag. To produce more stable estimates, sample covariances of different control runs were averaged together. In addition, following DelSole and Tippett (2009a), we truncate the sum in (A1) to 20 years and apply a Parzen window to the time-lagged covariances. The results are not sensitive to the truncation level as long as it is a small fraction of the total length of 4200.

To quantify time scale, we use the sample estimate of the integral time scale (1)

$$T_2 = 1 + 2\sum_{\tau=1}^{20} w_\tau \hat{\rho}_\tau^2, \tag{A2}$$

where $\hat{\rho}_\tau$ is the sample autocorrelation function of the time series, τ is time lag, in years, and w_τ are coefficients for the Parzen window. The sample integral time scale of the first 15 components in each control run are shown in fig. 11. The T_2 value of the leading component in each forced run is tabulated in table 2. For observation-based estimates, only results after 1900 are used– uncertainties prior to 1900 are deemed too large to allow useful estimates (though the results are not very different if data prior to 1900 is used.)

The horizontal line near the bottom of fig. 11 shows the 5% significance level, which was estimated by Monte Carlo methods as follows. An appropriate null hypothesis for our test is that the data is drawn from a white noise process when sampled every *two* years. It would be inappropriate to assume white noise for annual sampling because the ocean surface is highly correlated on monthly time scales and this correlation translates into a correlation between annual averages. Consistent with this, assuming white noise for annual sampling leads to

all components being statistically significant– i.e., all components are distinguishable from white noise. Because APT is invariant to nonsingular linear transformation, the process can be assumed to be white in space without loss of generality. Accordingly, we generate a 40 x 2100 data matrix by drawing independent random numbers from a normal distribution with zero mean and unit variance. Components that maximize APT were then determined. For each component, a corresponding 2100-year time series was derived. The integral time scale in each 150-year chunk was determined for each component, yielding 14 T_2-values per component. This procedure was repeated 100 times to generate 14 x 40 x 100 T_2-values. The upper five percentile of the 14 x 100 T_2-values of each of the 40 components was then determined. The leading 15 values are plotted in fig. 11 as the horizontal line. Note that the T_2 threshold values were doubled since the time step in the Monte Carlo method was two years. Values below this line can be interpreted as statistically indistinguishable from white noise (when sampled every two years). The figure shows that the first six components have statistically significant time scales. The leading component, called the Internal Multidecadal Pattern (IMP), has statistically significant T_2 values in all control runs.

d. Forced-to-Unforced Discriminants

The expected pattern of response to climate forcing is determined by discriminant analysis. This method assumes that the variability in the forced run can be modeled as internal noise plus an independently varying "signal" (i.e., the response to climate forcing). Under this assumption, the variance due to external forcing and internal variability are additive and hence the variance in the forced runs should be larger than in the unforced runs. There-

fore, we seek the pattern that maximizes the ratio of the variance in the forced runs to the variance in the unforced runs. This optimization problem is standard (Schneider and Held 2001; DelSole and Tippett 2007) and leads to the eigenvector problem

$$\mathbf{\Sigma}_f \mathbf{q} = \lambda \mathbf{\Sigma}_c \mathbf{q}, \tag{A3}$$

where $\mathbf{\Sigma}_f$ and $\mathbf{\Sigma}_c$ are covariance matrices for the forced and control runs, respectively, averaged over all models, $\mathbf{q}$ is the desired projection vector, and λ is an eigenvalue giving the variance ratio. The covariance matrices describe spatial covariability only– no time lag information is used in the covariance matrices. The variance in each control run is measured with respect to the 300-year mean of the control run, while variance in the forced run is measured with respect to the 1901-1950 mean of the respective forced run. Eigenvectors are ordered by decreasing eigenvalue, in which case the first maximizes the variance ratio, the second maximizes the variance ratio subject to being uncorrelated with the first, and so on. As in APT analysis, the data are projected onto the leading 40 EOFs of the multi-model ensemble. The leading eigenvector will be called the "forced-to-unforced discriminant."

e. Separating Forced and Unforced Components in Observations

To separate the forced and unforced components in observations, we represent the observed annual-mean temperature anomalies $T(x, y, t)$ as a combination of three components: the IMP pattern $p_I(x, y)$, a forced pattern $p_F(x, y)$, and noise ϵ:

$$T(x, y, t) = p_I(x, y)\beta_I(t) + p_F(x, y)\beta_F(t) + \epsilon(x, y, t), \tag{A4}$$

where β_I and β_F are time-varying amplitudes for the IMP and forced components, respectively, and ϵ is a random term that varies in space and time. The amplitudes β_I and β_F are determined using generalized least squares, which extends ordinary least squares to the case of dependent noise. This procedure is now recognized to be equivalent to fingerprinting (Allen and Tett 1999). To do this, we rewrite (A4) in matrix form as

$$\mathbf{T}_t = \mathbf{P}_t \boldsymbol{\beta}_t + \boldsymbol{\epsilon}_t, \tag{A5}$$

where $\mathbf{T}_t$ is an M-dimensional vector giving the observed temperature anomaly at time t, $\mathbf{P}_t$ is an $M \times 2$ matrix whose two columns are the patterns p_I and p_F, and the other terms are interpreted in an obvious manner. Although the patterns p_I and p_F do not change in time, their representation on the observation grid after missing values are masked out does. If $\boldsymbol{\epsilon}$ has covariance matrix $\sigma_\epsilon^2 \boldsymbol{\Sigma}_\epsilon$, then the generalized least squares estimate of $\boldsymbol{\beta}_t$ is

$$\hat{\boldsymbol{\beta}}_t = \left(\mathbf{P}_t^T \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{P}_t\right)^{-1} \mathbf{P}_t^T \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{T}_t, \tag{A6}$$

where the hat symbol denotes an estimated quantity. The estimated standard error of the i'th element of $\boldsymbol{\beta}_t$, denoted $\hat{se}(\boldsymbol{\beta})_i$, is

$$\hat{se}(\boldsymbol{\beta})_i = \sqrt{\hat{\sigma}_\epsilon^2 \left(\left(\mathbf{P}_t^T \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{P}_t\right)^{-1}\right)_{ii}} \tag{A7}$$

where $\hat{\sigma}_\epsilon^2$ is a sample estimate of σ_ϵ^2 given by

$$\hat{\sigma}_\epsilon^2 = \frac{1}{M-2} \left(\mathbf{T}_t - \mathbf{P}_t \hat{\boldsymbol{\beta}}_t\right)^T \boldsymbol{\Sigma}_\epsilon^{-1} \left(\mathbf{T}_t - \mathbf{P}_t \hat{\boldsymbol{\beta}}_t\right). \tag{A8}$$

The noise covariance matrix $\boldsymbol{\Sigma}_\epsilon$ is estimated by the average sample covariance matrix of the individual control runs. Although $\boldsymbol{\Sigma}_\epsilon$ does not change with time, the pattern of observations with non-missing values depends on time, and hence the covariance matrix formed by

extracting the matrix elements corresponding to the observed grid points depends on time. The condition number of the noise covariance matrix was found to depend significantly with the observation network, with large condition numbers occurring in the pre-1900 era. To avoid unstable estimates, the inverse covariance matrix was approximated using only the leading 20 eigenvectors of the noise covariance matrix in each year. A technical point is that all fields need to be area weighted before computing the eigenvectors, to ensure that the pseudo-inverse is taken with respect to an area weighted norm.

The amplitudes of the forced and IMP components for 10-40 EOFs are shown in fig. 12. As can be seen, the amplitudes are nearly independent of EOF truncation for years after 1900. The results are more sensitive for 6 or fewer EOFs. The greater uncertainty in the pre-1900 period is presumably due to greater missing data that occurs during this period.

f. Missing Observations

The standard error estimate in (A7) does not depend on missing data and hence does not account for uncertainty due to missing data. To estimate this uncertainty, we adopt the following resampling procedure. First, a period with reasonably complete observations was identified. We found that the 25-year period 1981-2005 had no more than 20 missing values in any single year out of the 688 grid cells used in the calculations. This period is accordingly identified as the "data-rich period." Then, for each year, the amplitudes of the patterns were computed for a year in the data-rich period for two different networks, namely the network of observations in the data-rich period, and the network of observations in the year in question. The difference between these two estimates gives an estimate of the uncertainty due to the

missing data. Repeating this for all 25 years in the data-rich period gives 25 error estimates from which the variance can be estimated. This variance is then added to the variance of the regression estimate (i.e., added to the square of equation(A7)) to obtain an estimate of the total uncertainty variance due to both finite sample size and missing observations.

The above approach to dealing with missing observations is not claimed to be optimal. Other methods, such as the regularized imputation method proposed by Schneider (2001), may produce more accurate estimates by exploiting space-time correlations in the data.

g. The Ensemble Mean IMP and its Time Scale

The ensemble mean IMP shown in fig. 6 is calculated as the average IMP of 36 forced runs from the leading 14 models listed in table 2. The error bars are computed as $\pm\sigma/\sqrt{36}$, where σ is the standard deviation of the IMP in the forced runs at each year.

To estimate the time scale of the ensemble mean IMP, we fit the ensemble mean IMP from the forced runs to the first order autoregressive model

$$x_t = \phi x_{t-1} + k + \epsilon_t, \tag{A9}$$

where x_t is the ensemble mean IMP, ϵ_t is independent random error with zero mean and constant variance, and k is a constant. Standard regression techniques give the estimates $\phi = 0.87$ and $k = -0.07$. Following Leith (1973), the effective time scale is computed as

$$T = \sum_{\tau=0}^{\infty} \rho(\tau) = \sum_{\tau=0}^{\infty} \phi^{\tau} = 1/(1-\phi), \tag{A10}$$

where $\rho(\tau) = \phi^{\tau}$ is the autocorrelation function of the process. Substituting $\phi = 0.87$ gives the effective time scale $T \approx 7.7$ years.

h. Distribution of Trends for a Stochastic Process

The trend of a stationary process z_t for $t \in [0, T]$ is obtained by fitting the equation

$$z_t = \beta (t - t_0) + \epsilon_t, \tag{A11}$$

where

$$t_0 = \frac{1}{T+1} \sum_{t=0}^{T} t. \tag{A12}$$

The least squares estimate of the trend parameter β is

$$\hat{\beta} = \frac{\sum_{t=0}^{T} z_t (t - t_0)}{\sum_{t=0}^{T} (t - t_0)^2}. \tag{A13}$$

Since the sum of $t - t_0$ over $t \in [0, T]$ vanishes, the mean of $\hat{\beta}$ vanishes. The variance of $\hat{\beta}$ is

$$var[\hat{\beta}] = \frac{\sum_{t=0}^{T} \sum_{t'=0}^{T} (t - t_0)(t - t_0) E[z_t z_{t'}]}{\left(\sum_{t=0}^{T} (t - t_0)^2 \right)^2} \tag{A14}$$

$$= \frac{\sum_{\tau=-T}^{T} \sigma_z^2 \sum_{t'=0}^{T-|\tau|} \rho_\tau (t - t_0)(t - t_0 + |\tau|)]}{\left(\sum_{t=0}^{T} (t - t_0)^2 \right)^2} \tag{A15}$$

$$= \left(\frac{\sigma_z^2}{\sum_{t=0}^{T} (t - t_0)^2} \right) \left(1 + 2 \sum_{\tau=1}^{T} \rho_\tau c_\tau \right) \tag{A16}$$

where

$$c_\tau = \frac{\sum_{t=0}^{T-\tau} (t - t_0)(t - t_0 + \tau)}{\sum_{t=0}^{T} (t - t_0)^2} \tag{A17}$$

The first term in parentheses in (A16) is the variance of the T-year trend for a white noise process with variance σ_z. The second term in parentheses is an "inflation factor" that accounts for autocorrelation in the time series.

To estimate the variance of trends for the IMP, we fit the IMP to a autoregressive model and use the resulting model to estimate the autocorrelation ρ_τ. The autocorrelation function

for the best fit AR1 and AR2 models turn out to be virtually indistinguishable. The best fit AR1 model of the form (A9), where x_t is substituted for the IMP, over the period 1900-2008 is found to be $\phi = 0.806$. The best fit AR1 models for 1900-1945 and 1946-2008 give statistically indistinguishable results. The standard deviation of the spatially averaged IMP is $\sigma_{IMP} = 0.0872$, which was substituted into (A16). The 95% confidence interval for the the trend is then $\pm 1.96\sqrt{var[\hat{\beta}]}$, which is plotted in fig. 8 as a function of trend period T.

REFERENCES

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

Barnett, T. P., 1984: Long-term trends in surface temperature over the oceans. *Mon. Wea. Rev.*, **112**, 303–312.

Bindoff, N. L., et al., 2007: Observations: Oceanic climate change and sea level. *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor, and H. L. Miller, Eds., Cambridge University Press, chap. 5, 385–432.

Bjerknes, J., 1964: Atlantic air-sea interaction. *Advances in Geophysics*, Academic Press, 1–82.

DelSole, T. and J. Shukla, 2010: Model fidelity versus skill in seasonal forecasting. *J. Climate*, in press, available from http://journals.ametsoc.org/doi/pdf/10.1175/2010JCLI3164.1.

DelSole, T. and M. K. Tippett, 2007: Predictability: Recent insights from information theory. *Rev. Geophys.*, doi:10.1029/2006RG000202.

DelSole, T. and M. K. Tippett, 2009a: Average predictability time: Part I. Theory. *J. Atmos. Sci*, **66**, 1172–1187.

DelSole, T. and M. K. Tippett, 2009b: Average predictability time: Part II: Seamless diagnosis of predictability on multiple time scales. *J. Atmos. Sci.*, **66**, 1188–1204.

Delworth, T., S. Manabe, and R. J. Stouffer, 1993: Interdecadal variations of the thermohaline circulation in a coupled ocean-atmosphere model. *J. Climate*, **6**, 1993–2011.

Delworth, T. L. and M. E. Mann, 2000: Observed and simulated multidecadal variability in the Northern Hemisphere. *Clim. Dyn.*, **16**, 661–676.

Forster, P., et al., 2007: Changes in atmospheric constituents and in radiative forcing. *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor, and H. L. Miller, Eds., Cambridge University Press, chap. 2, 129–234.

Ghil, M., et al., 2002: Advanced spectral methods for climatic time series. *Rev. Geophys.*, **40**, 3.1–3.41.

Hasselmann, K., 1997: Multi-pattern fingerprint method for detection and attribution of climate change. *Clim. Dyn.*, **13**, 601 – 611.

Hasselmann, K. F., 1979: On the signal-to-noise problem in atmospheric response studies. *Meteorology of the tropical ocean*, D. B. Shaw, Ed., Royal Meteorological Society, 251–259.

Hegerl, G. C., et al., 2007: Understanding and attributing climate change. *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, S. Solomon, D. Qin,

M. Manning, Z. Chen, M. Marquis, K. Averyt, M. Tignor, and H. Miller, Eds., Cambridge University Press, 663–745.

Huang, N. E. and Z. Wu, 2008: A review on Hilbert-Huang transform: method and its applications to geophysical studies. *Rev. Geophys.*, **46**, RG2006, doi:10.1029/2007RG000 228.

Huntingford, C., P. A. Stott, M. R. Allen, and F. H. Lambert, 2006: Incorporating model uncertainty into attribution of observed temperature change. *Geophys. Res. Lett.*, **33**, L05 710.

Jolliffe, I. T. and D. B. Stephenson, (Eds.) , 2003: *Forecast verification: A Practitioner's Guide in Atmospheric Science*. Wiley-Interscience, 254 pp.

Knight, J. R., R. J. Allan, C. K. Folland, and M. Vellinga, 2005: A signature of persistent natural thermohaline cirulation cycles in observed climate. *Geophys. Res. Lett.*, **32 (L20708)**, doi:10.1029/2005GL024 233.

Knutson, T. R., et al., 2010: Tropical cyclones and climate change. *Nature Geosci.*, **3**, 157–163, DOI:10.1038/NGEO779.

Kushnir, Y., 1994: Interdecadal variations in the North Atlantic sea surface temperature and associated atmospheric conditions. *J. Climate*, **7**, 141–157.

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

Leith, C. E., 1973: The standard error of time-average estimates of climate means. *J. Appl. Meteor.*, **12**, 1066–1069.

Lemke, P., et al., 2007: Observations: Changes in snow, ice and frozen ground. *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor, and H. L. Miller, Eds., Cambridge University Press, chap. 4, 337–383.

Rayner, N. A., P. Brohan, D. E. Parker, C. F. Folland, J. J. Kennedy, M. Vanicek, T. Answell, and S. Tett, 2006: Improved analyses of changes and uncertainties in sea surface temperature measured in situ since the mid-nineteenth century: the HadSST2 data set. *J. Climate*, **19**, 446–469.

Sato, M., J. E. Hansen, M. P. McCormick, and J. B. Pollack, 1993: Stratospheric aerosol optimal depths 1850-1990. *J. Geophys. Res.*, **98(D12)**, 22 987–22 994.

Schlesinger, M. E. and N. Ramankutty, 1994: An oscillation in the global climate system of period 65-70 years. *Nature*, **367**, 723–726.

Schneider, T., 2001: Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. *J. Climate*, **14**, 853–871.

Schneider, T. and I. M. Held, 2001: Discriminants of twentieth-century changes in earth surface temperatures. *J. Climate*, **14**, 249–254.

Stone, D. A., M. R. Allen, and P. A. Stott, 2007: A multi-model update on the detection and attribution of global surface warming. *J. Climate*, **14**, 3551–3565.

Straus, D., J. Shukla, D. Paolino, S. Schubert, M. Suarez, P. Pegion, and A. Kumar,

2003: Predictability of seasonal mean atmospheric circulation during Autumn, Winter, and Spring. *J. Climate*, **16**, 3629–3649.

Thompson, D. W. J., J. J. Kennedy, J. M. Wallace, and P. D. Jones, 2008: A large discontinuity in the mid-twentieth century in observed global-mean surface temperature. *Nature*, **453**, 646–649.

Ting, M., Y. Kushnir, R. Seager, and C. Li, 2009: Forced and internal twentieth-century SST in the North Atlantic. *J. Climate*, **22**, 1469–1481.

Trenberth, K. E., et al., 2007: Observations: Surface and atmospheric climate change. *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor, and H. L. Miller, Eds., Cambridge University Press, chap. 3, 235–335.

Venzke, S., M. R. Allen, R. T. Sutton, and D. P. Rowell, 1999: The atmospheric response over the North Atlantic to decadal changes in sea surface temperature. *J. Climate*, **12**, 2562–2584.

Zhang, R., T. L. Delworth, and I. M. Held, 2007: Can the Atlantic Ocean drive the observed multidecadal variability in the Northern Hemisphere mean temperature? *Geophys. Res. Lett.*, **34 (L02709)**, doi:10.1029/2006GL028 683.

Zwiers, F. W. and X. Zhang, 2003: Toward regional scale climate change detection. *J. Climate*, **16**, 793–797.

List of Tables

Period	Data Set	Trend (K/decade)
1946 - 1977	Total Observations	0.0378 ± 0.0372
1946 - 1977	Forced + IMP	0.0486 ± 0.0291
1946 - 1977	Forced Only	0.112 ± 0.0355
1977 - 2008	Total Observations	0.145 ± 0.029
1977 - 2008	Forced + IMP	0.166 ± 0.0319
1977 - 2008	Forced Only	0.122 ± 0.0314
1946 - 2008	Total Observations	0.0909 ± 0.0137
1946 - 2008	Forced + IMP	0.096 ± 0.0133
1946 - 2008	Forced Only	0.103 ± 0.012

TABLE 1. Trends in annual-mean, spatially-averaged SST for different reconstructions and periods. "Total observations" refers to well-observed HadSST2, "Forced + IMP" refers to best fit SST using forced component plus IMP, and "Forced only" refers to best fit SST using the forced component only. Trends are expressed in K/decade plus-minus the 95% confidence interval.

	model	NAT	sims	T_2 (years)	variance
1	cccma_cgcm3_1	-	5	7 ± 2	0.44 ± 0.14
2	cccma_cgcm3_1_t63	-	1	5.6	0.88
3	cnrm_cm3	-	1	6.8	0.94
4	csiro_mk3_0	-	3	7 ± 1.6	3.09 ± 1.1
5	csiro_mk3_5	-	1	6.3	0.82
6	gfdl_cm2_0	Y	3	5.2 ± 2.2	1.17 ± 0.37
7	gfdl_cm2_1	Y	5	5.9 ± 1.2	1.39 ± 0.44
8	inmcm3_0	-	1	5	0.8
9	ipsl_cm4	-	1	5.3	0.99
10	miroc3_2_medres	Y	3	8.5 ± 0.55	0.43 ± 0.00
11	miub_echo_g	Y	3	4.5 ± 2	1.31 ± 0.87
12	mri_cgcm2_3_2a	Y	5	6.6 ± 3.7	0.51 ± 0.1
13	ncar_ccsm3_0	Y	2	11.2 ± 6.8	1.57 ± 0.49
14	ukmo_hadcm3	-	2	7.7 ± 2.4	0.82 ± 0.00
15	bccr_bcm2_0	-	1	9.7	0.81
16	giss_aom	-	2	3.5 ± 0.32	0.64 ± 0.1
17	giss_model_e_h	Y	5	6.6 ± 2	1.7 ± 1.2
18	giss_model_e_r	Y	9	7.9 ± 7.9	0.73 ± 0.2
19	ingv_echam4	-	1	3.1	0.31
20	miroc3_2_hires	-	1	6	0.5
21	mpi_echam5	-	3	7.8 ± 2.4	0.67 ± 0.14
22	ncar_pcm1	Y	3	6.3 ± 0.55	0.95 ± 0.26
23	ukmo_hadgem1	Y	2	3.5 ± 0.45	0.8 ± 0.1
24	hadsst2		1	6.3	1.65

TABLE 2. The number of ensemble members ("sims"), indicator of whether the model contained natural forcing ("NAT"; "Y" if yes, "-" if not), integral time scale T_2 (in years), and variance of the Internal Multidecadal Pattern estimated from each model of the forced runs (first 23 rows), and observations (last row). For consistency, all statistics are computed using only the last 100 years of the available time series. The first 14 rows specify the models used to optimize for the forced and internal patterns. For model forced runs with more than one ensemble member, the mean value and standard deviation of T_2 and variance are indicated in the table.

List of Figures

4 Generalized least squares estimates of the amplitude of the forced component (top) and the IMP (bottom) when the forced component is determined from the forced-to-unforced discriminant. The shading indicates twice the standard error (top) and the standard error (bottom) of the estimates as estimated from standard regression theory plus a contribution due to missing data. The blue curve in the upper panel indicates the ensemble mean time series of the forced-to-unforced discriminant in the forced runs. The blue dashed lines in the top panel indicate years of the five most significant volcanic eruptions after 1850, in terms of the change in visible optimal depth as estimated by Sato et al. (1993); specifically, Krakatoa in 1883, Santa Maria in 1902, Mt. Agung in 1961, El Chichon in 1982, and Mt. Pinatubo in 1991. The red curve in the bottom panel shows the annual average Atlantic Multidecadal Oscillation Index after rescaling. 45

5 The squared autocorrelation function of the IMP in each model control run as a function of time lag. The 5% significance level of the autocorrelation, for a sample size of 300, is indicated by the thick horizontal dashed line. 46

6 Ensemble mean time series of the Internal Multidecadal Pattern in the forced runs. The ensemble mean is an average over all forced runs used to calculate the forced-to-unforced discriminant pattern. The error bars show the 95% confidence interval for the ensemble mean. The red dashed line indicates zero amplitude. 47

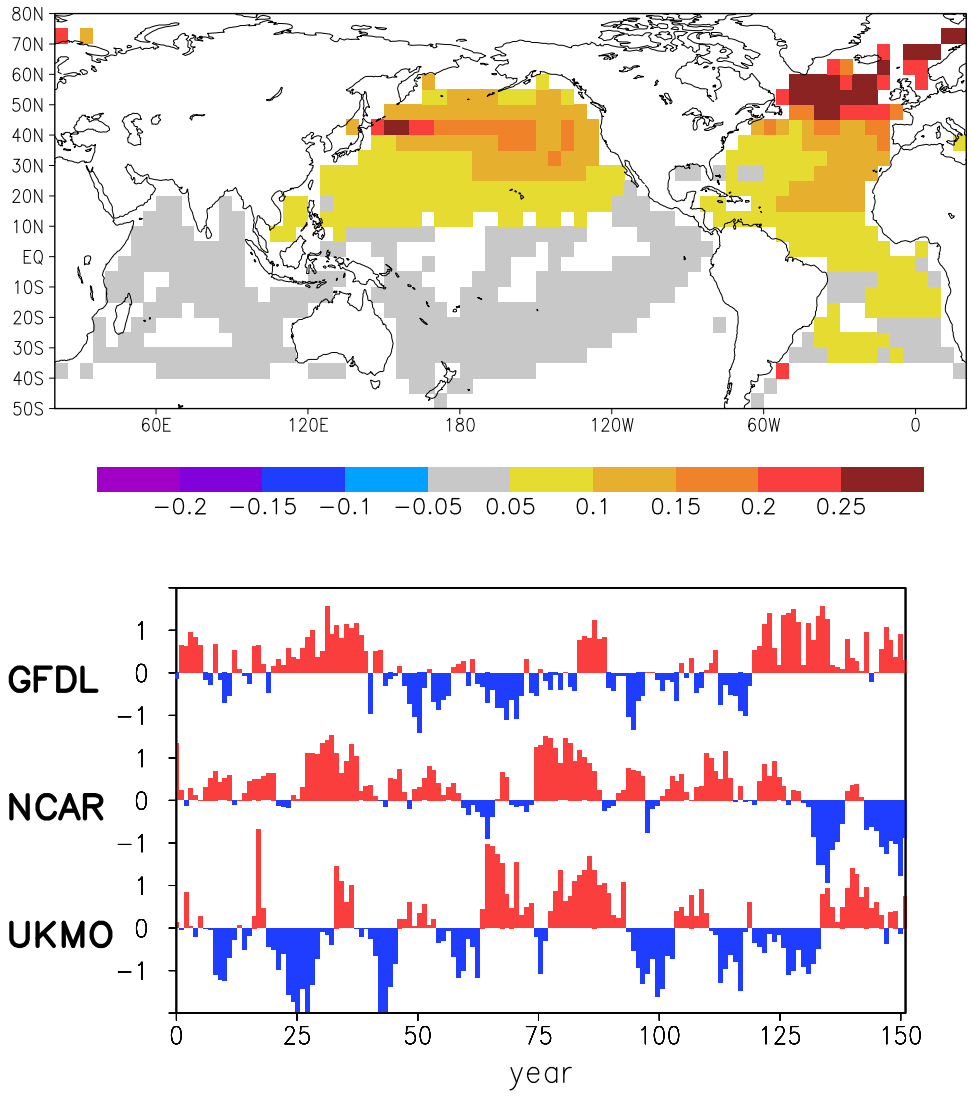

FIG. 1. The component that maximizes the average predictability time of sea surface temperature in 14 climate models run with fixed forcing (i.e., "control runs"). The top panel shows the spatial structure of the component. This component is called the Internal Multidecadal Pattern, or IMP. Ocean points with no shading indicate regions that were omitted from the maximization (i.e., "masked out") because of insufficient data in the corresponding observational data set. The bottom panel shows the time series of this component in three representative control runs: ukmo_hadcm3 ("UKMO"), ncar_ccsm3_0 ("NCAR"), and gfdl_cm2_1 ("GFDL").

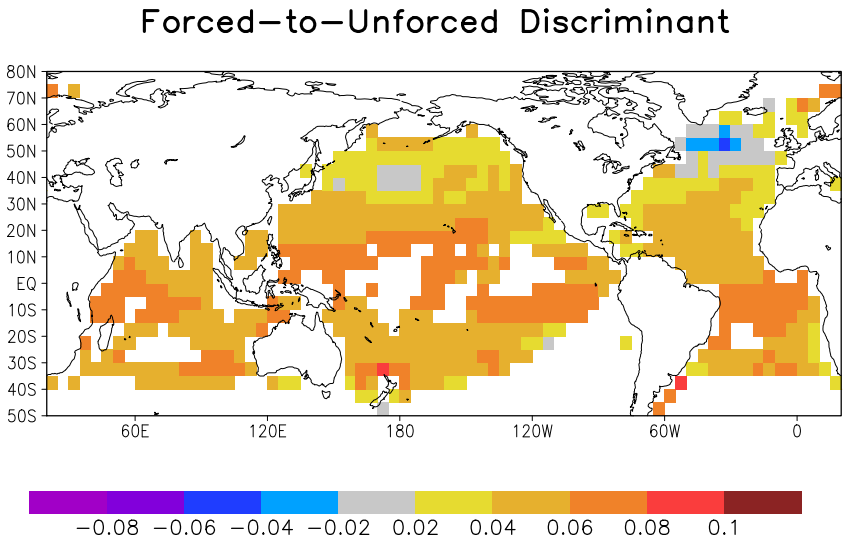

FIG. 2. The pattern of the expected response to climate forcing, obtained by maximizing the ratio of variances between the forced and control simulations.

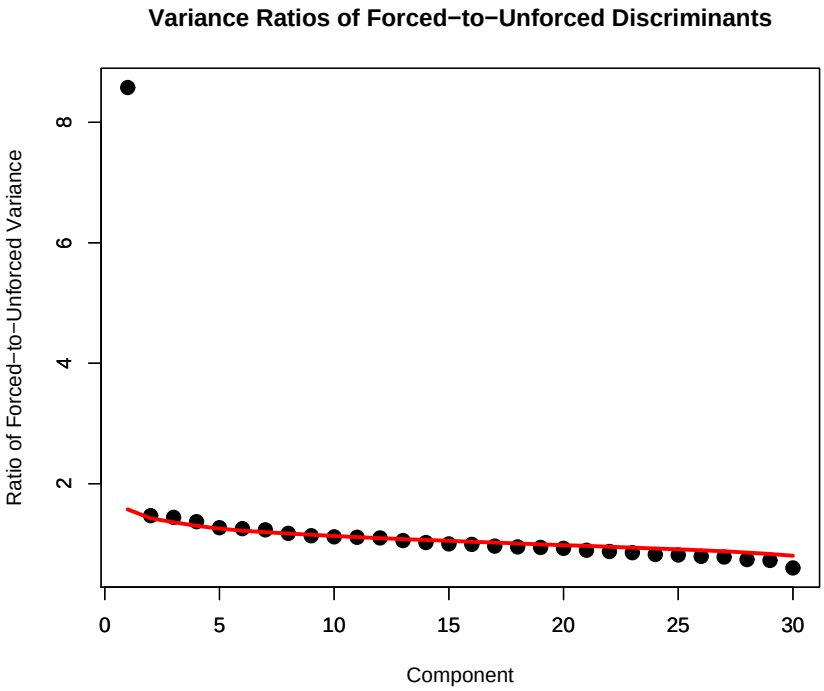

FIG. 3. Optimized ratio of forced variance to unforced variance, as determined by discriminant analysis of the leading 30 multimodel EOFs. The red curve shows the 5% significance threshold determined from bootstrap methods.

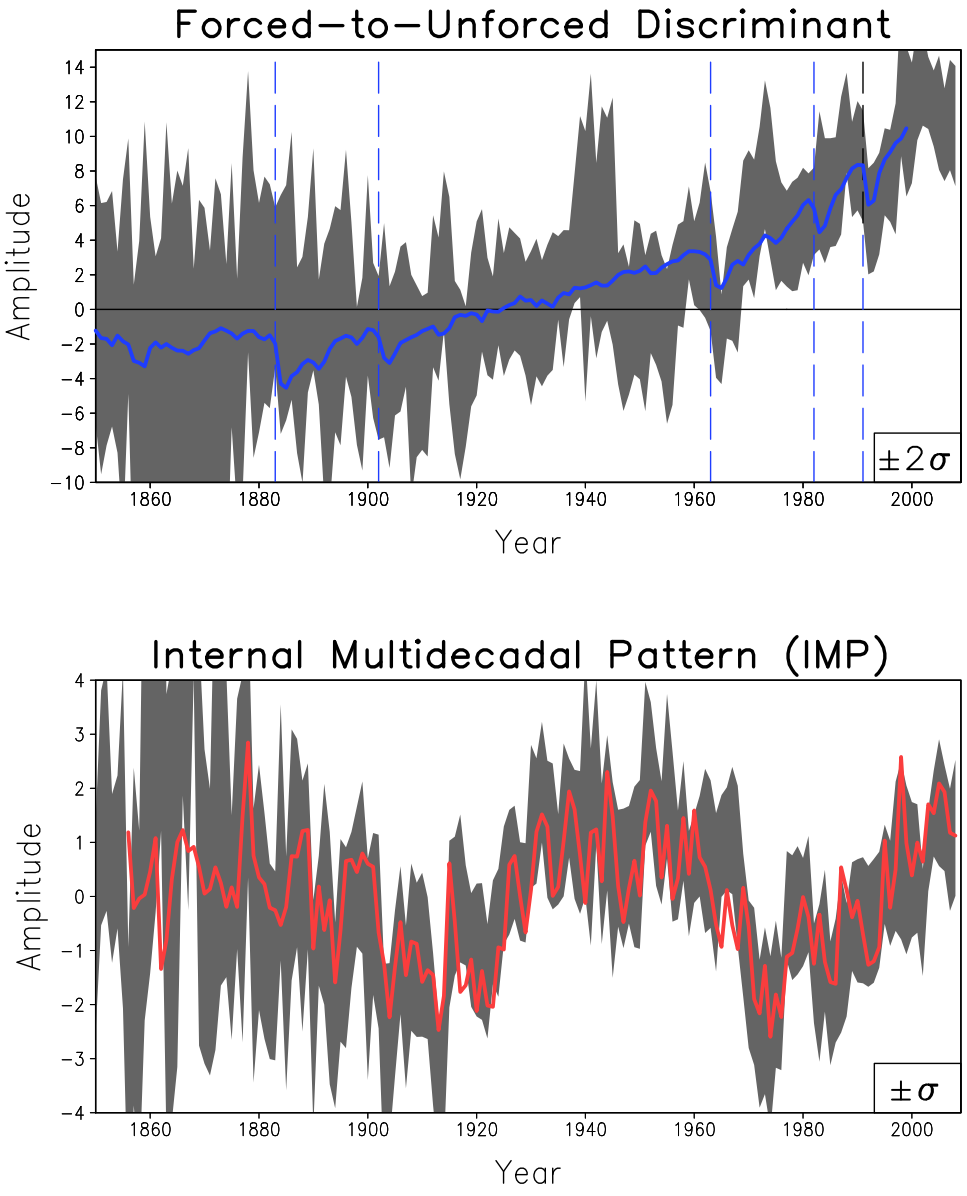

FIG. 4. Generalized least squares estimates of the amplitude of the forced component (top) and the IMP (bottom) when the forced component is determined from the forced-to-unforced discriminant. The shading indicates twice the standard error (top) and the standard error (bottom) of the estimates as estimated from standard regression theory plus a contribution due to missing data. The blue curve in the upper panel indicates the ensemble mean time series of the forced-to-unforced discriminant in the forced runs. The blue dashed lines in the top panel indicate years of the five most significant volcanic eruptions after 1850, in terms of the change in visible optimal depth as estimated by Sato et al. (1993); specifically, Krakatoa in 1883, Santa Maria in 1902, Mt. Agung in 1961, El Chichon in 1982, and Mt. Pinatubo in 1991. The red curve in the bottom panel shows the annual average Atlantic Multidecadal Oscillation Index after rescaling.

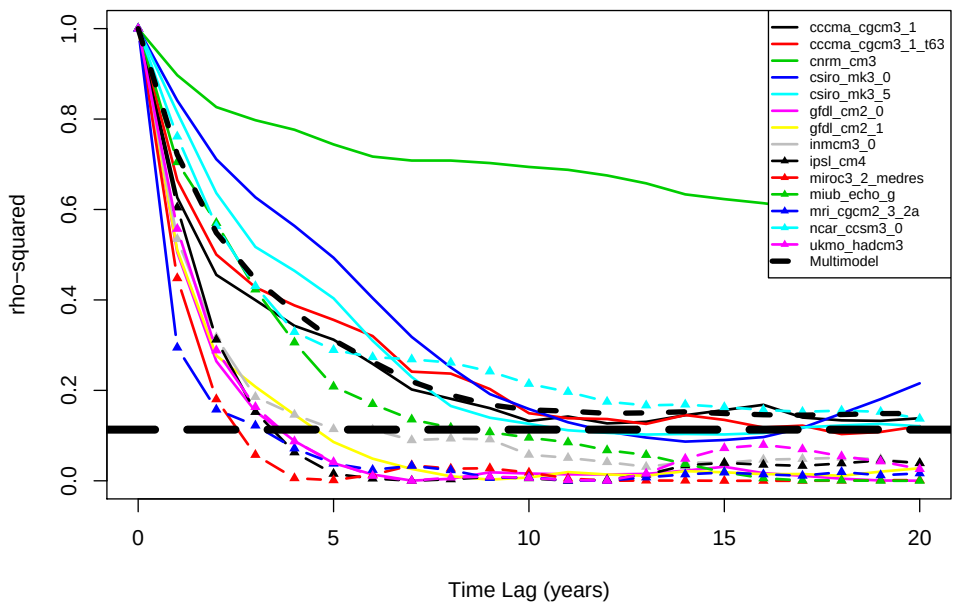

FIG. 5. The squared autocorrelation function of the IMP in each model control run as a function of time lag. The 5% significance level of the autocorrelation, for a sample size of 300, is indicated by the thick horizontal dashed line.

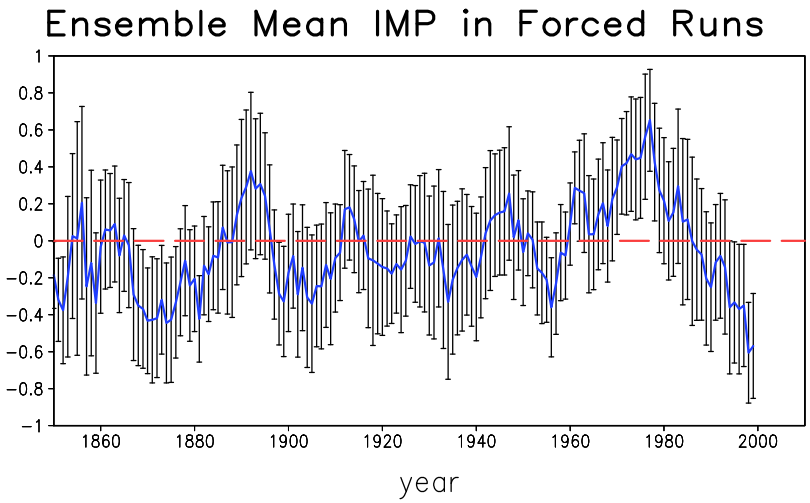

FIG. 6. Ensemble mean time series of the Internal Multidecadal Pattern in the forced runs. The ensemble mean is an average over all forced runs used to calculate the forced-to-unforced discriminant pattern. The error bars show the 95% confidence interval for the ensemble mean. The red dashed line indicates zero amplitude.

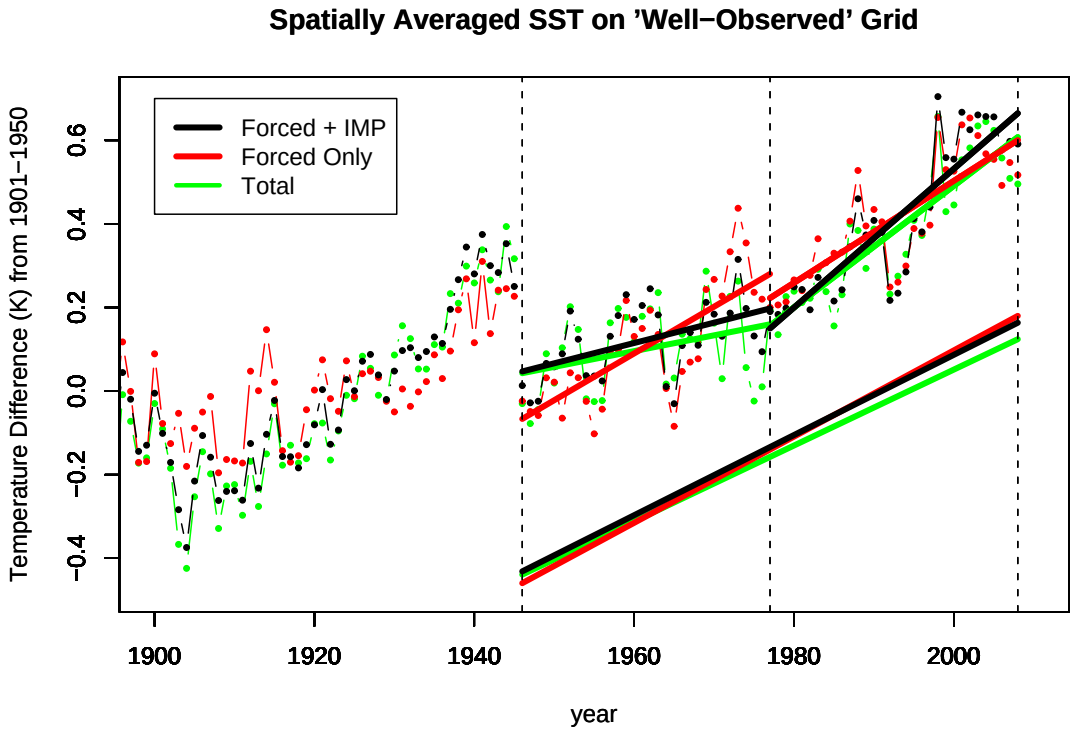

FIG. 7. The spatially averaged sea surface temperature on the "well-observed grid" for observations (green dots), as reconstructed by the sum of the forced component and IMP (black dots), and as reconstructed by the forced component only (red dots). The amplitudes are relative to the 1901-1950 mean amplitude. The best fit linear trends for the periods 1946-1977, 1977-2008, and 1946-2008 are shown as solid lines, with the trend for the last period offset by -0.4K for clarity. The actual trend values are given in table 1.

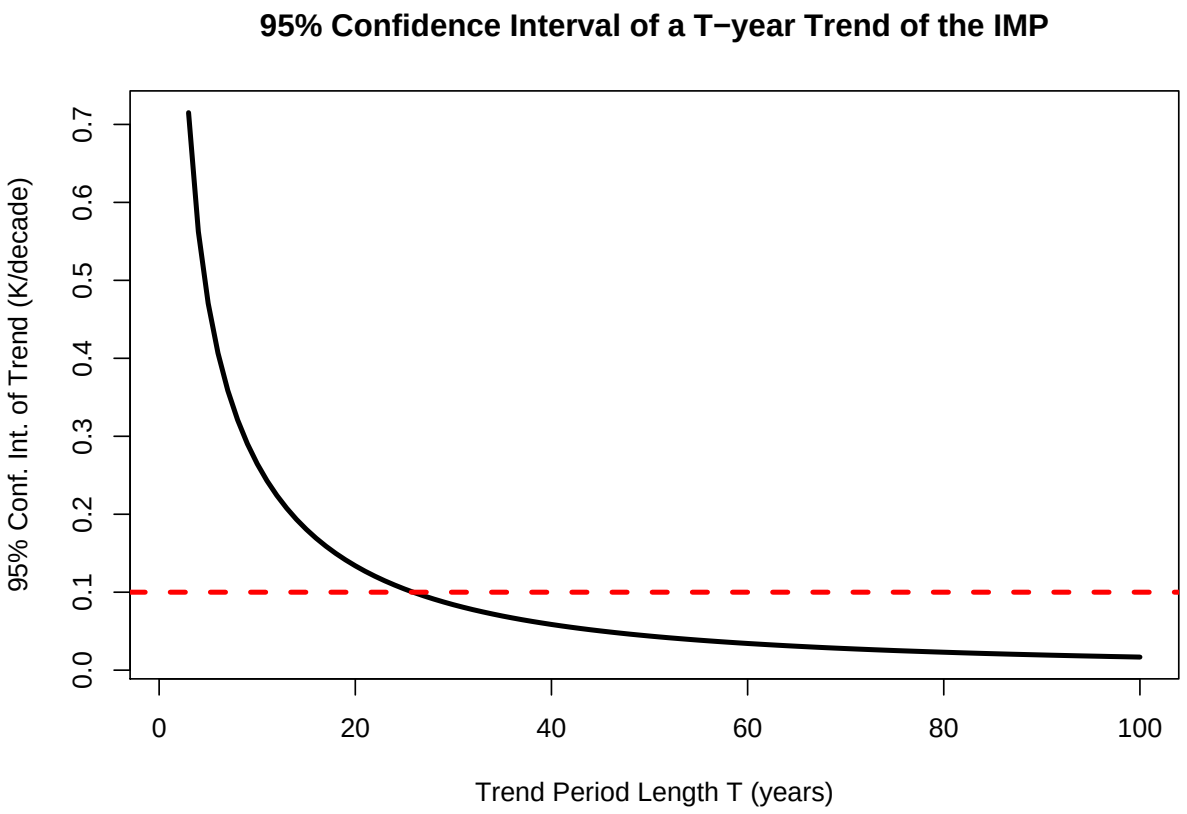

FIG. 8. Estimated 95% confidence interval for the trend due to the IMP as a function of trend period length. The estimate is based on a first order autoregressive model fit to the IMP with AR-parameter 0.806. The horizontal dashed line indicates the estimated trend due to anthropogenic and natural forcing.

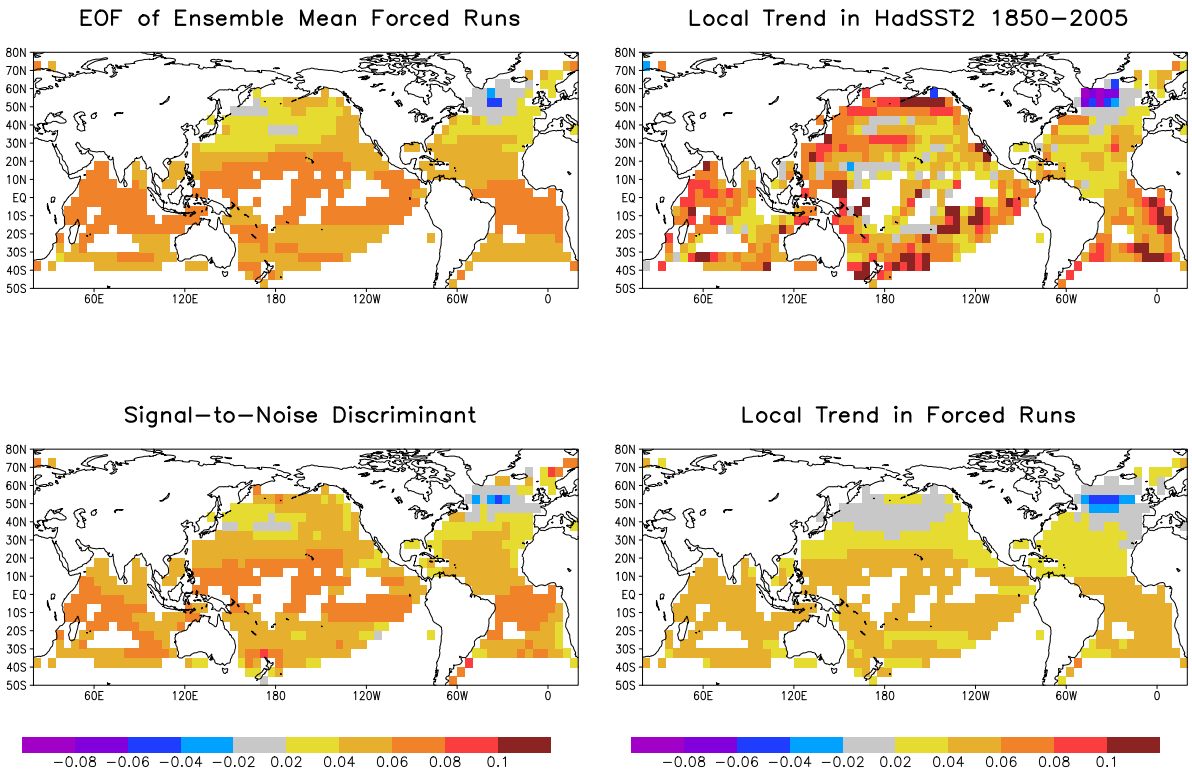

FIG. 9. Four alternative estimates of the forced pattern. The different methods are indicated in the title of each panel.

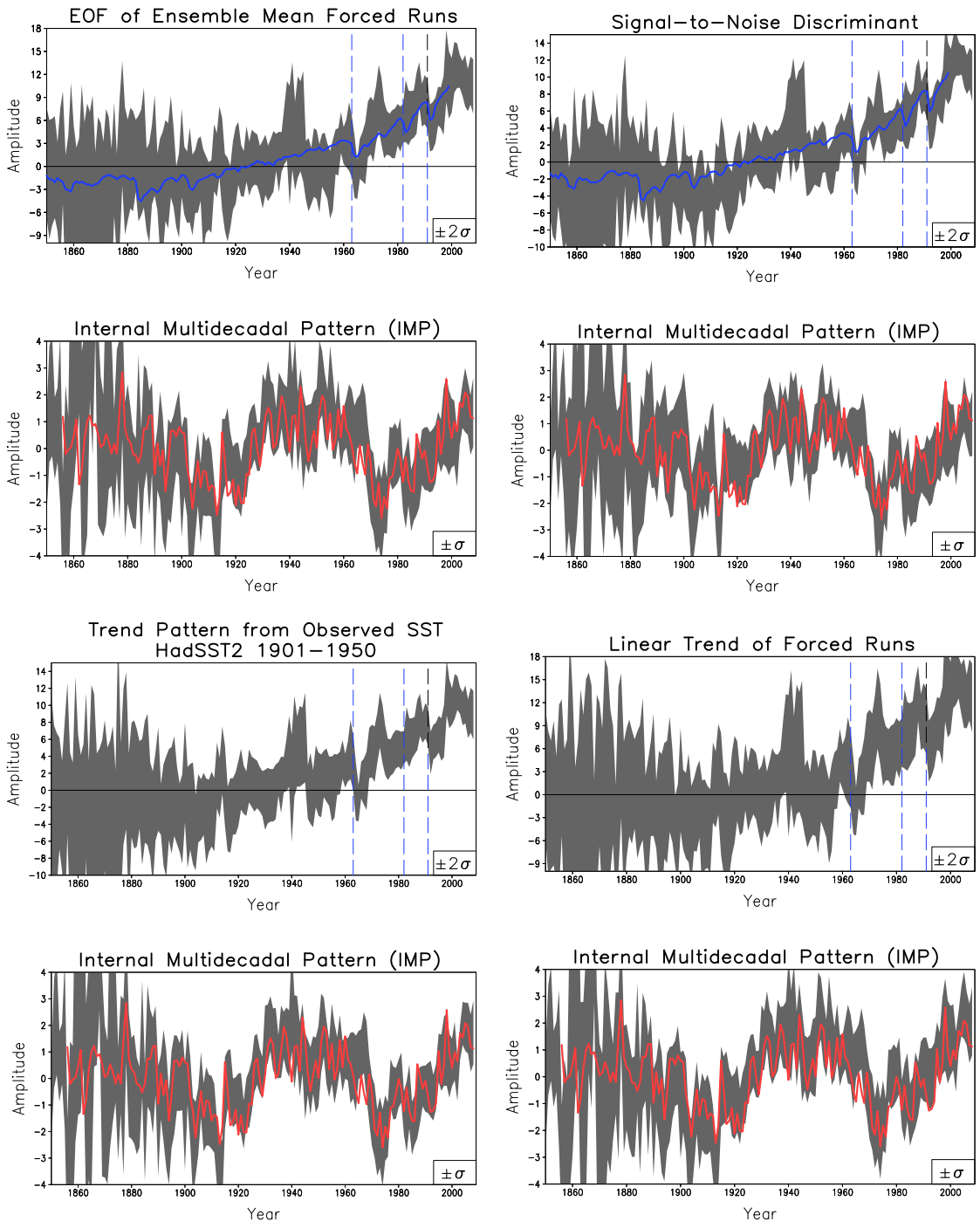

FIG. 10. Time series for the forced pattern and IMP for the four different estimates of the forced pattern shown in fig. 9. The respective forced pattern is indicated in the title of each panel, and the corresponding IMP is indicated in the panel directly below the time series for the forced pattern.

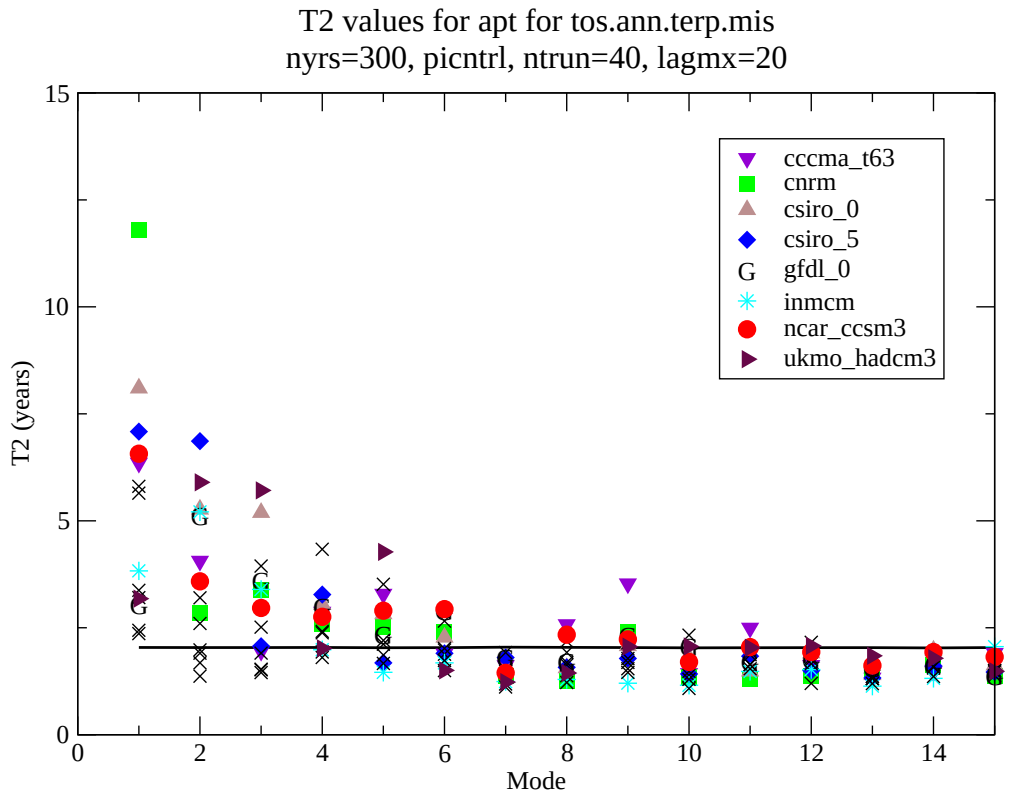

FIG. 11. T_2 time scales of components that maximize the average predictability time of the 14-control runs. The integral time scale of selected models are displayed by different symbols as indicated in the legend. The solid horizontal line is the 5% significance level of the integral time scale.

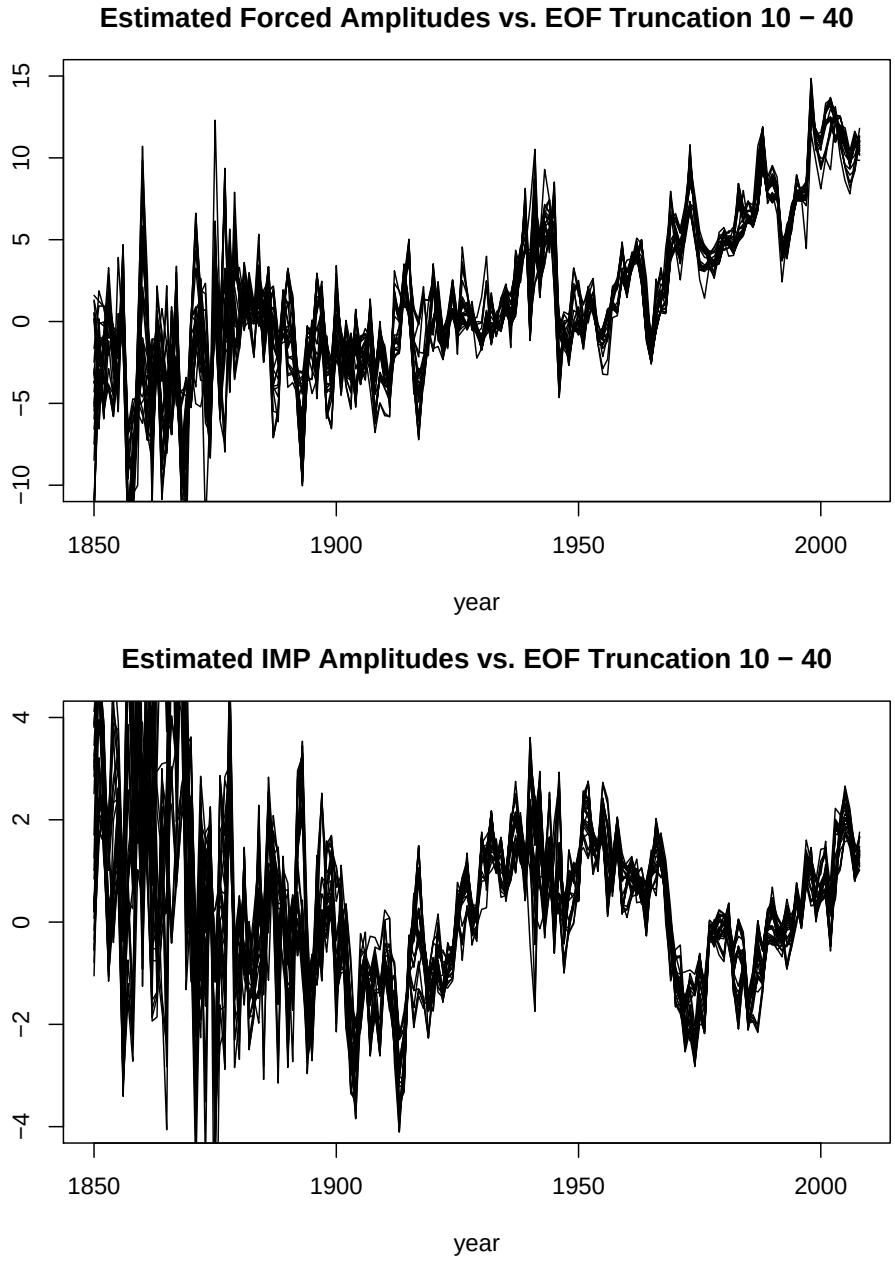

FIG. 12. Amplitudes of the forced component (top panel) and the IMP (bottom panel) in the HadSST2 data set using 10-40 leading EOFs. The result of each EOF truncation is shown as a separate curve.