Separating Internal Variability from the Externally Forced Climate Response

LEELA M. FRANKCOMBE AND MATTHEW H. ENGLAND

ARC Centre of Excellence for Climate System Science, and Climate Change Research Centre, University of New South Wales, Sydney, New South Wales, Australia

MICHAEL E. MANN

Department of Meteorology, and Earth and Environmental Systems Institute, The Pennsylvania State University, University Park, Pennsylvania

BYRON A. STEINMAN

Large Lakes Observatory, and Department of Earth and Environmental Sciences, University of Minnesota Duluth, Duluth, Minnesota

(Manuscript received 21 January 2015, in final form 17 June 2015)

ABSTRACT

Separating low-frequency internal variability of the climate system from the forced signal is essential to better understand anthropogenic climate change as well as internal climate variability. Here both synthetic time series and the historical simulations from phase 5 of CMIP (CMIP5) are used to examine several methods of performing this separation. Linear detrending, as is commonly used in studies of low-frequency climate variability, is found to introduce large biases in both amplitude and phase of the estimated internal variability. Using estimates of the forced signal obtained from ensembles of climate simulations can reduce these biases, particularly when the forced signal is scaled to match the historical time series of each ensemble member. These so-called scaling methods also provide estimates of model sensitivities to different types of external forcing. Applying the methods to observations of the Atlantic multidecadal oscillation leads to different estimates of the phase of this mode of variability in recent decades.

1. Introduction

Internally generated natural variability is an important part of the climate system. Although the longest-term, largest-scale climate trends are dominated by external forcing, internal variability plays a vital role at shorter time scales and at smaller spatial scales. An example is the recent slowdown in global surface warming, which has led to heightened scrutiny of the role played by both forced and internal climate variability on decadal to multidecadal time scales. Among the outstanding underlying issues is how best to separate internal variability from the forced climate signal.

For the actual climate, we have only one realization of the internal variability and it is nontrivial to extract it from the available data. Schurer et al. (2013) used proxy reconstructions and model simulations to estimate the contributions of internal variability and external forcing over the last millennium. Estimating the forced signal during the historical era is complicated by the short length of the observational record and the challenge this creates in isolating low-frequency, multidecadal, and longer-term internal variability (Frankcombe et al. 2015). In addition, the dominant influence on climate in the most recent period is anthropogenic forcing, including greenhouse gases (GHGs), tropospheric aerosols, and ozone-depleting substances, each of which must separately be taken into account. One recent body of research, for example, has sought to ascertain how much of the mid-twentieth-century temperature variability is due to anthropogenic aerosols and how much is due to internal variability (Booth et al. 2012; Zhang et al. 2013). Mann et al. (2014) used observations to investigate the effect of biases caused by the incorrect partition of observed Northern Hemisphere temperatures into forced

Corresponding author address: Leela M. Frankcombe, Climate Change Research Centre, Level 4 Mathews Bldg., University of New South Wales, Sydney, NSW 2052, Australia.
E-mail: l.frankcombe@unsw.edu.au

DOI: 10.1175/JCLI-D-15-0069.1

and internal components. Steinman et al. (2015) extended that work to study the relative contributions of the North Atlantic and North Pacific to the observed internal variability of the Northern Hemisphere.

In this paper we compare various methods for separating the forced signal from the background of internal variability and examine the biases that may result from the different methods. We focus on the specific example of multidecadal North Atlantic sea surface temperature (SST) variability, but the results have broader implications for the problem of separating forced and internal climate variability.

Enhanced variability on multidecadal time scales centered in the North Atlantic has been found in modern observational climate data (Folland et al. 1984, 1986; Kushnir 1994; Mann and Park 1994; Delworth and Mann 2000) and in long-term climate proxy data (e.g., Mann et al. 1995; Delworth and Mann 2000). Such variability is also generated in a range of models from idealized ocean models to full GCMs (Delworth et al. 1993, 1997; Huck et al. 1999; Knight et al. 2005; Parker et al. 2007; Ting et al. 2011; Zhang and Wang 2013). The variability has been named the Atlantic multidecadal oscillation (AMO; Kerr 2000) or, alternatively, Atlantic multidecadal variability (AMV) since it is unclear whether it truly constitutes a narrowband oscillatory climate signal. In this study, we do not attempt to address the mechanisms causing the variability; we instead focus on North Atlantic SST variability as a case study in the application of competing statistical approaches to separating internal and external variability.

The rest of this paper is divided as follows: We first describe the data used in the study (section 2) and then describe the various competing methods for separating forced and internal variability (section 3). The methods are tested on synthetic data, where the true internal and external signals are known (section 4), and then applied to CMIP5 historical simulations (section 5) and observational data (section 6). We then discuss the results of our analyses (section 7) and finally summarize with our conclusions (section 8).

2. Data

One often-used measure of AMV is the smoothed and linearly detrended average of North Atlantic SSTs (e.g., Sutton and Hodson 2003). We calculate an index of North Atlantic variability by averaging SST over the region 0°–60°N, 5°–75°W but do not detrend the series, for reasons that will become clear later in the discussion. We will call this raw index the North Atlantic SST index (NASSTI). Estimates of the internal variability obtained from the NASSTI using the methods tested here are referred to as Atlantic multidecadal oscillation indices (AMOI), since they are approximations of AMO/AMV variability. We use the historical runs from phase 5 of the Coupled Model Intercomparison Project (CMIP5; Taylor et al. 2012), employing the 145-yr (1861–2005) interval spanned by nearly all ensemble members. Simulations that do not span the full interval are excluded, as are models in which the raw NASSTI time series does not display significant multidecadal variability. Two idealized historical scenarios—Hist_{GHG} (in which only well-mixed greenhouse gas forcing is applied) and Hist_{Nat} (natural forcings only, including solar variability and volcanoes)—are also used. The CMIP5 models used are listed in Table 1. For comparison to observations we use SST from HadISST (Rayner et al. 2003) between 1870 and 2005. Smoothed time series are calculated using a 40-yr adaptive low-pass filter (Mann 2008).

3. Methods

Of the many methods used to separate the forced signal and the internal variability, the most common is the ''detrended'' approach, where a linear trend is subtracted from the signal (e.g., Zhang and Wang 2013). This method has the advantage of being extremely simple and, in the absence of any better estimates of the forced signal, may also be useful as a first approximation. The external forcing is not linear in time, however. For this reason, the detrending procedure has been shown to bias the amplitude and phase of the estimated internal variability (Mann and Emanuel 2006; Mann et al. 2014). Biases in the estimated phase will in turn bias estimates of AMO periodicity.

An alternative method, referred to as the ''differenced'' method, employs a large ensemble of climate simulations. Each individual ensemble member responds to the external forcing applied to the model, but it also contains a realization of internal variability. If the ensemble members are initialized so as to be independent of each other, then they will each contain a different realization of the internal variability. Averaging over a large number of these ensemble members will average out the internal variability so that the signal remaining is the model response to the external forcing. Subtracting this model-mean response from each ensemble member gives the internal variability. This method has the advantage that it does not make prior assumptions about the model response to external forcing. The method does, however, rely on each member of the ensemble having the same response to the external forcing, which is not necessarily the case. The strength of a model's response to external forcing is

TABLE 1. CMIP5 models used. Some models list more than one control run length, indicating that various sections of control runs were available. (Expansions of acronyms are available at http://www.ametsoc.org/PubsAcronymList.)

Model name	Length of control run (yr)	Historical	$Hist_{Nat}$	$Hist_{GHG}$
BCC_CSM1.1	500	3	1	1
BCC_CSM1.1(m)	400	3	—	—
BNU-ESM	559	1	—	—
CanESM2	996	5	5	5
CMCC-CESM	277	1	—	—
CMCC-CM	330	1	—	—
CMCC-CMS	500	1	—	—
CNRM-CM5	850	10	6	6
CNRM-CM5.2	410	1	—	—
ACCESS1.0	500	2	—	—
ACCESS1.3	500	3	2	2
CSIRO Mk3.6.0	500	10	5	5
FIO-ESM	800	3	—	—
EC-EARTH	451	13	—	—
INM-CM4.0	500	1	—	—
IPSL-CM5A-LR	1000	6	3	5
IPSL-CM5A-MR	300	3	3	3
IPSL-CM5B-LR	300	1	—	—
FGOALS-g2	700	4	3	1
MIROC-ESM	680	3	3	3
MIROC-ESM-CHEM	255	1	1	1
MIROC5	—	5	—	—
HadCM3	—	10	10	—
HadGEM2-CC	240	1	—	—
HadGEM2-ES	577	5	4	4
MPI-ESM-LR	1000	3	—	—
MPI-ESM-MR	1000	3	—	—
MPI-ESM-P	1155	2	—	—
MRI-CGCM3	500	5	1	1
MRI-ESM1	—	1	—	—
GISS-E2-H	1470 + 531	15	5	5
GISS-E2-H-CC	251	1	—	—
GISS-E2-R	251 + 1200 + 531 + 531 + 1163	25	10	5
GISS-E2-R-CC	250	1	—	—
CCSM4	1051 + 156	6	4	3
NorESM1-M	500	3	1	1
NorESM1-ME	251	1	—	—
GFDL CM2.1	—	10	10	—
GFDL CM3	500	5	3	3
GFDL-ESM2G	500	1	—	—
GFDL-ESM2M	500	1	1	1
CESM1(BGC)[a]	500	1	—	—
CESM1(CAM5)	319	3	3	3
CESM1(CAM5, FV2)[b]	—	4	2	1
CESM1(FASTCHEM)	222	3	—	—
CESM1(WACCM)	200	1	—	—
CSIRO Mk3L.1.2[c]	1000	2	—	—
Total		194	66	59

[a] BGC indicates biogeochemistry.
[b] FV2 indicates finite-volume dynamical core with 2° model output.
[c] A combination of the low-resolution atmospheric component of CSIRO Mk3 and the ocean component of CSIRO Mk2.

represented by the equilibrium climate sensitivity (ECS), which is the equilibrium change in annual global-mean surface temperature after a doubling of the atmospheric CO_2 concentration relative to preindustrial levels. The CMIP5 models have equilibrium climate sensitivities of between 2.1° and 4.7°C (Flato et al. 2013). Even for an ensemble of realizations from a single climate model, the estimates of climate sensitivity derived from a single ensemble member may differ from the true model sensitivity because of the noise introduced by

internal variability (Huber et al. 2014). Furthermore, in the case of a multimodel ensemble, each individual model will have a different climate sensitivity altogether. The multimodel mean (MMM) represents an average across models that both overestimate (i.e., high-sensitivity models) and underestimate (i.e., low-sensitivity models) the forced response. The MMM will therefore overestimate the magnitude of the forced response for models with low sensitivity and underestimate it for models with high sensitivity. The differenced method thus potentially introduces a bias when used to estimate the internal variability of the various models. Although small during the earlier part of the historical record when the amplitude of the forced signal is modest, the bias becomes significant toward the end of the historical period and increasingly dominates over the signal of internal variability in any future projections.

One method to mitigate this bias, the ''scaling'' method, is described by Steinman et al. (2015). In this method the multimodel mean of the CMIP5 historical all-forcing ensemble is taken to be the best estimate of the climate response to external forcing and is then scaled to match the climate sensitivity of each individual ensemble member. For the test case described here the multimodel mean of the NASSTI is linearly regressed on to the NASSTI time series of each ensemble member from the CMIP5 historical all-forcing ensemble to obtain an estimate of the forced signal:

$$R_1(t) = \beta_c + \beta \mathrm{MMM}_{\mathrm{all}}(t), \quad (1)$$

where β_c is a constant, β is the scaling factor, and $\mathrm{MMM}_{\mathrm{all}}$ is the multimodel mean of the NASSTI from the CMIP5 all-forcing ensemble. The regression coefficient β is a measure of the relative climate sensitivity of each ensemble member compared to $\mathrm{MMM}_{\mathrm{all}}$ and is thus model dependent. The component of the time series of each ensemble member not explained by the scaled multimodel mean is taken as an estimate of the internal variability in the North Atlantic N_1 and is recovered by subtracting R_1, the estimate of each ensemble member's forced response, from H, the time series of each ensemble member from the historical simulation:

$$N_1(t) = H(t) - R_1(t). \quad (2)$$

This method, which we term the ''single factor scaling'' method, results in much better estimation of phase and amplitude of low-frequency variability than the detrending and differencing methods (Steinman et al. 2015). It, too, however, is not completely free of potential biases. Consider that external forcing during the historical period has contributions from both greenhouse gases and aerosols (both anthropogenic and volcanic) and that different models may have different amplitude responses to the different types of forcing. Indeed, different models may have different specifications and implementations of the various forcing components. The single factor scaling method, however, uses a single regression coefficient to account for all external forcing. While the method performs well over the historical period (Steinman et al. 2015), application of the method to future projections, which contain an increasingly large contribution from one particular forcing (anthropogenic greenhouse gases), could result in biases at the ends of the time series.

In addition to the single factor scaling method we test two modified scaling methods where two or three scaling factors are used. While in the single factor scaling method the (single) scaling factor represents the combined model response to all external forcings, in the modified scaling approaches different scaling factors are used to represent the model responses to different types of external forcing—in effect the different efficacies of the different forcings. For the modified scaling method using two scaling factors, estimates of the two factors for each time series are calculated by multilinear regression on the NASSTI time series of each ensemble member:

$$R_2(t) = \gamma_c + \gamma_{\mathrm{GHG}} \mathrm{MMM}_{\mathrm{GHG}}(t) + \gamma_{\mathrm{Nat}} \mathrm{MMM}_{\mathrm{Nat}}(t), \quad (3)$$

where γ_c is a constant and γ_{GHG} and γ_{Nat} are the estimated scaling factors. The first scaling factor represents the model response to GHG forcing, while the second represents the model response to natural forcing, such as volcanic aerosols and solar variability. The estimates of the GHG and natural responses are obtained from the multimodel means of the $\mathrm{Hist}_{\mathrm{GHG}}$ and $\mathrm{Hist}_{\mathrm{Nat}}$ simulations of CMIP5 ($\mathrm{MMM}_{\mathrm{GHG}}$ and $\mathrm{MMM}_{\mathrm{Nat}}$, respectively). The resulting estimate of the forced response is used to recover an estimate of the internal variability as follows:

$$N_2(t) = H(t) - R_2(t). \quad (4)$$

In addition to GHG and natural forcings there are also other forcings included in the all-forcing experiments that should be taken into account (anthropogenic aerosols and ozone being the most important in the context of North Atlantic multidecadal variability), but these cannot be robustly included because of the limited number of ensemble members that performed these individual forcing experiments. If sufficient simulations of the various other forcings were available, then scaling factors representing them could be included, in addition to the scaling factors representing GHG and natural forcings. As an estimate of these unrepresented forcings we include a third scaling factor $\mathrm{MMM}_{\mathrm{rest}}$, which is the multimodel mean of the variability that remains

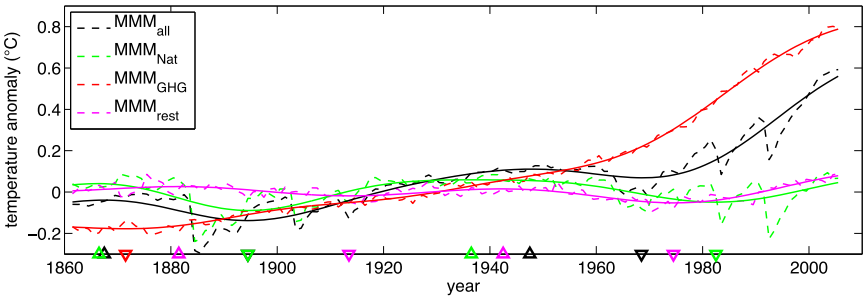

FIG. 1. Multimodel means of the NASSTI in the all-forcing ensemble (black), the natural forcings ensemble (green), the GHG forcing ensemble (red), and the remainder after natural forcing and GHG forcing are removed from the all-forcing ensemble (magenta). Annual data are shown by the dashed lines while data smoothed with a 40-yr low-pass filter are shown by the solid lines. Upward (downward) pointing triangles on the x axis indicate the position of maxima (minima) of the four smoothed time series.

unexplained after regressing $\mathrm{MMM}_{\mathrm{GHG}}$ and $\mathrm{MMM}_{\mathrm{Nat}}$ on $\mathrm{MMM}_{\mathrm{all}}$:

$$\mathrm{MMM}_{\mathrm{rest}}(t) = \mathrm{MMM}_{\mathrm{all}}(t) - \varepsilon_{\mathrm{GHG}}\mathrm{MMM}_{\mathrm{GHG}}(t) + \varepsilon_{\mathrm{Nat}}\mathrm{MMM}_{\mathrm{Nat}}(t). \quad (5)$$

The three factor scaling method is calculated as follows:

$$R_3(t) = \delta_c + \delta_{\mathrm{GHG}}\mathrm{MMM}_{\mathrm{GHG}}(t) + \delta_{\mathrm{Nat}}\mathrm{MMM}_{\mathrm{Nat}}(t) + \delta_{\mathrm{rest}}\mathrm{MMM}_{\mathrm{rest}}(t) \quad \text{and} \quad (6)$$

$$N_3(t) = H(t) - R_3(t), \quad (7)$$

where δ_c is a constant, and δ_{GHG}, δ_{Nat}, and δ_{rest} are the estimated scaling factors for GHG forcing, natural forcing, and residual forcing, respectively. The various MMMs are shown in Fig. 1. Note that forcings included in the all-forcing historical simulations but not in $\mathrm{Hist}_{\mathrm{GHG}}$ or $\mathrm{Hist}_{\mathrm{Nat}}$ may have, in addition to the forced signal represented by $\mathrm{MMM}_{\mathrm{rest}}$, additional projections onto $\mathrm{MMM}_{\mathrm{GHG}}$ and $\mathrm{MMM}_{\mathrm{Nat}}$ such that δ_{GHG} and δ_{Nat} (and indeed γ_{GHG} and γ_{Nat} in the two scaling factor method) represent sensitivities to combinations of forcings.

These scaling methods are analogous to the methods of optimal fingerprinting used in detection and attribution studies (Allen and Tett 1999; Allen and Stott 2003). The difference here is that we use a single time series rather than spatial patterns and focus on extracting the natural variability rather than the forced signal. The three scaling methods were tested with both ordinary least squares regression (as used by Steinman et al. 2015) and total least squares regression (Allen and Stott 2003); no significant differences were found between the two regression methods.

The multiensemble, multimodel mean of the CMIP5 historical runs is used as the estimate of the forced signal for the differenced and single factor scaling approaches. Each ensemble member from each model is given equal weight in the mean, which can lead to biasing toward models that contribute a large ensemble to the CMIP5 archive. However, averaging the ensemble of each model to get a model mean and then averaging all the model means to get a multimodel mean, as is sometimes done to account for differing ensemble sizes, results in the internal variability of the members of large ensembles being averaged out before they can contribute to the multimodel mean. This method implicitly assumes that internal variability is negligible and, in the presence of the nonnegligible internal variability that is of interest in this study, results in a bias toward the models that contribute fewer ensemble members to the archive (since each of the few ensemble members effectively receives a larger weight in the multimodel averaging process). In choosing to calculate the forced signal as a multiensemble mean we are implicitly assuming that all the ensemble members, from all the models, are drawn from the same distribution (i.e., that all the models perform equally). The limitations of this assumption will be investigated later.

4. Analysis of the various methods using synthetic data

To test the various methods in an idealized situation where the true internal variability is known, we construct synthetic AMOI time series using the null hypothesis that the variability is due to red noise. Each synthetic time series of internal variability N is a 145-yr-long time series of red noise (the same length as the CMIP5 historical runs), scaled by the average autocorrelation and amplitude of the CMIP5 historical runs. Three independent, random scaling factors (drawn from the uniform distribution between 0.2 and 2) are used—the first representing the response to GHG forcing

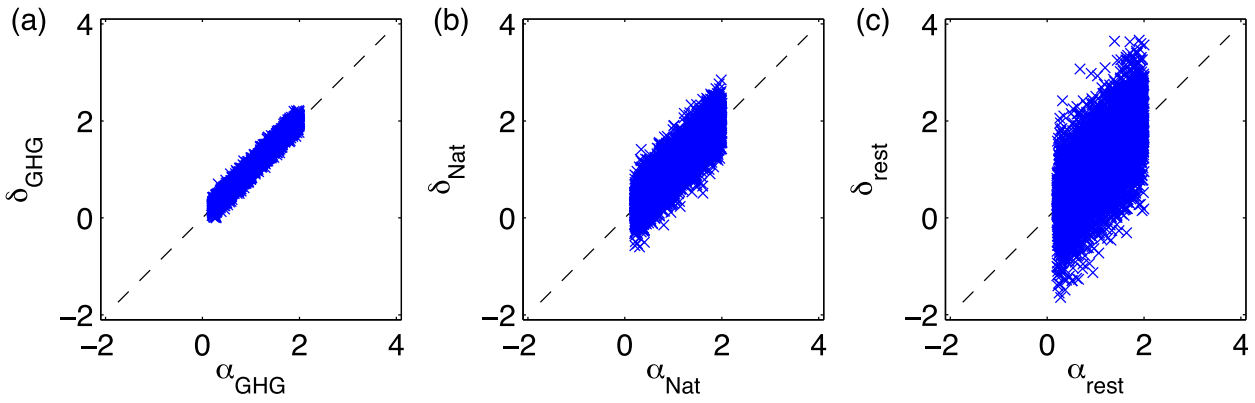

FIG. 2. Scatterplots of the known scaling coefficients compared to the estimates made using the three scale factor method for (a) α_{GHG} vs δ_{GHG}, (b) α_{Nat} vs δ_{Nat}, and (c) α_{rest} vs δ_{rest} for 5000 synthetic time series.

(α_{GHG}), the second the response to natural forcing (α_{Nat}), and the third the response to any other forcings (α_{rest}). The independence of the scaling factors is shown in section 5 to be valid for the CMIP5 models; thus we use that assumption for the synthetic time series here. The synthetic historical time series were constructed by adding forced variability to the natural variability as follows:

$$H(t) = N(t) + \alpha_{GHG}\mathrm{MMM}_{GHG}(t) + \alpha_{Nat}\mathrm{MMM}_{Nat}(t) + \alpha_{rest}\mathrm{MMM}_{rest}(t), \quad (8)$$

where H is the synthetic historical time series; N is the synthetic time series of internal variability; and MMM_{GHG}, MMM_{Nat}, and MMM_{rest} are the multimodel means of the NASSTI time series representing GHG, natural, and residual forcings from CMIP5, respectively (as shown in Fig. 1). An ensemble of 5000 such time series was constructed.

The five methods to remove the forced signal are then applied to the synthetic data to find N_{est}, the estimated internal variability. The accuracy of the methods can be judged by comparing the estimated internal variability N_{est} to the true time series N using a variety of metrics:

1) comparing the estimated scaling factors (β, γ, and δ) to the known ones (α),
2) calculating error as a function of time,
3) finding minima and maxima of the estimated time series compared to the known ones (to find the bias in phase introduced by each method), and
4) calculating the amplitudes of the estimated time series compared to the known ones (to find the bias in amplitude introduced by each method).

This gives us a basis for comparison for the CMIP5 models, for which the true time series of internal variability are not known.

Figure 2 shows scatterplots of the estimated scaling factors compared to the known scaling factors for the three factor scaling method. In Fig. 2a we can see that the true GHG scaling factor α_{GHG} is well estimated by δ_{GHG}, the GHG scaling factor from the three factor method. This is also the case for the two factor scaling method, with α_{GHG} and γ_{GHG} being highly correlated. For the single factor scaling method the scaling factor β also correlates very well with α_{GHG}, while the correlation of β with α_{Nat} is small, although not negligible, with higher α_{Nat} on average corresponding to larger β for the same value of α_{GHG}. This indicates that it is the model sensitivity to GHG forcing which dominates over the sensitivity to natural forcing in the single scaling method. Figure 2b shows the accuracy with which α_{Nat} is estimated using the three factor scaling method. The accuracy is very similar for the two factor scaling method. The accuracy of estimation of α_{rest} is shown in Fig. 2c. This factor is the most difficult to estimate because MMM_{rest} varies on similar time scales to the internal variability, so the two may easily be mistaken for each other. The error in estimating α_{Nat} is smaller but arises from the same source since the natural forcing also contains variability on multidecadal time scales. The error in estimating α_{GHG} is the smallest of the three; therefore, sensitivity to GHG forcing should be the most robustly estimated parameter.

The error in each estimation can be calculated as a function of time:

$$\mathrm{Err}(t) = \sqrt{[N_{est}(t) - N(t)]^2}. \quad (9)$$

Figure 3a shows the mean error as a function of time for the synthetic time series for the five different methods. The raw NASSTI time series (gray lines) has errors that increase with time as the external forcing becomes increasingly dominant. The detrending method (blue

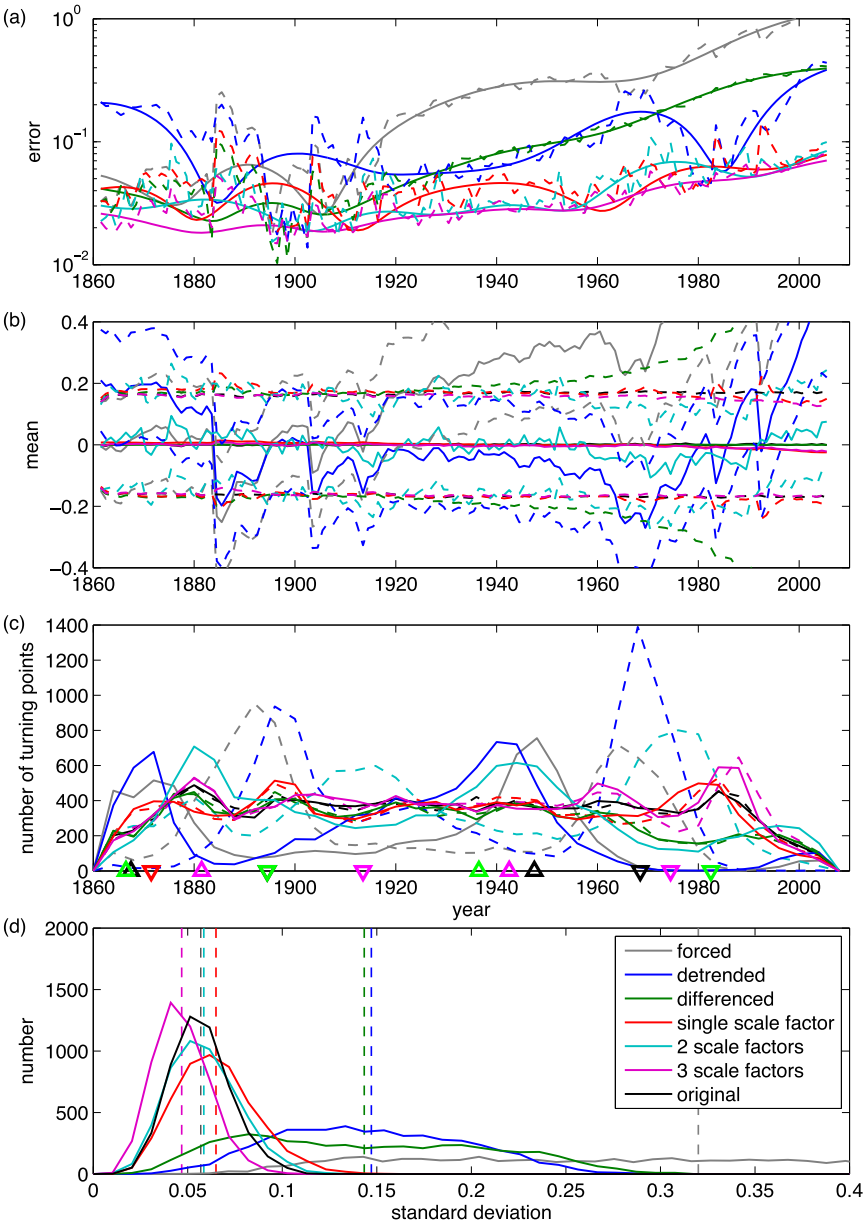

FIG. 3. Time series of (a) mean error as a function of time (dashed; annual mean, solid; after 40-yr smoothing; note the log scale on the *y* axis) and (b) mean (solid) and one standard deviation on either side of the mean (dashed) as a function of time of the 5000 synthetic time series for the five methods. (c) Distribution of turning points as a function of time (solid lines indicating maxima and dashed lines indicating minima), with triangles on the *x* axis indicating minima and maxima of the MMMs as in Fig. 1. (d) Distribution of standard deviations of the estimated variability for each method for the synthetic time series. Dashed vertical lines in (d) indicate the means of the distributions.

lines) has large errors through the whole time series, particularly at the beginning and end owing to the assumption that the trend is linear. Errors in the differencing method (green lines) increase toward the end of the time series as a result of the increasing influence of different models' climate sensitivities. The single factor scaling method (red lines) gives smaller errors than the detrending and differencing methods, especially toward the end of the time series, because the MMM is matched to the model climate sensitivity by the scaling. Errors at the beginning of the time series, however, are comparable to the differenced method because GHG forcing is

small and differing climate sensitivities of the models thus have a minor impact. Errors using the single scaling method increase during volcanic eruptions because the single scaling factor is more sensitive to the GHG response than the naturally forced response. Using two scale factors (light blue lines) reduces the error in the 1940s (when there was a peak in MMM_{Nat}; see Fig. 1) but not elsewhere, while the three factor scaling method (magenta lines) results in a general improvement over the other methods.

The means (solid curves) and standard deviations (dashed curves) of the time series of estimated internal variability are plotted in Fig. 3b as a function of time. By construction, as the number of time series increases, the mean of the true time series of internal variability approaches zero and the standard deviation approaches a constant. The accuracy of the various methods is assessed in comparison to this. This metric shows similar results to Fig. 3a and is included for comparison with the CMIP5 models, where the error cannot be directly calculated since the true time series of internal variability are not known.

The raw forced signal (gray) shows increasing deviation from the true time series (in black). The mean of the detrended time series (blue) shows anomalous behavior particularly at the beginning and end. The mean for the differenced case (green) is always zero by construction (since we are subtracting the mean, the sum of the remainders will be zero), while the standard deviation shows a large increase at the end of the run. The single factor scaling method (red) shows a slightly larger spread of amplitudes around the times of volcanic eruptions. The mean for the two factor scaling case (light blue) shows larger departures from zero than the other two scaling cases during several periods (associated with turning points of MMM_{rest}; see Fig. 1), indicating that the forced signal has not been completely removed. The three factor scaling method (magenta) generally shows the least spread, at times even having a lower standard deviation than the true time series. The reason for this reduction in amplitude will be discussed later. The discrepancies between the various estimates relative to the true time series all correspond to periods where the errors (in Fig. 3a) are the largest.

To show the bias in the phase of the internal variability estimated using the various methods, the turning points of the 40-yr smoothed time series are plotted in Fig. 3c. Unbiased time series should show a uniform distribution of both maxima (solid lines) and minima (dashed lines) throughout the historical period. The true time series (black), however, shows a decreasing number of both maxima and minima about 20 years from the beginning and end of the time series as a result of the edge effects of the 40-yr smoothing (which should therefore be common to all five methods). Both the raw forced time series (gray) and the detrended time series (blue) have a bias toward minima in the 1890s and 1970s with maxima in between, corresponding to turning points in MMM_{Nat} (marked on the x axis in Fig. 3c). Both methods also show very few maxima after the 1960s because of the increasing dominance of the anthropogenic warming signal, which is not correctly removed. For the same reason, the differencing method (green) also shows a decrease in the number of turning points toward the end of the time series, which is larger than the filtering-induced decrease. The single scaling method (red) does a much improved job of finding the maxima and minima, while the two factor scaling method (light blue) results in large numbers of maxima around 1880 and 1940 and minima in the 1910s and 1970s, coinciding with turning points of MMM_{rest}. The additional external forcing represented by MMM_{rest} is already implicitly included in MMM_{all}, which is used in the single scaling method, but it is not represented by either MMM_{GHG} or MMM_{Nat} used in the two factor scaling method, which explains why the single scale factor method outperforms the two scale factor method when estimating phase of the internal variability. Of the five methods, the three factor scaling method (magenta) comes the closest to reproducing the true distribution of phases.

The distribution of amplitudes of the 40-yr smoothed time series of the estimated internal variability is shown in Fig. 3d. Both detrending and differencing results in a large overestimation of the amplitude. The scaling methods all do a better job of estimating the amplitude, although the single factor scaling method overestimates the amplitude while the three factor scaling method underestimates it. In the single scaling method this is due to the sometimes incomplete removal of the natural forcing signal, which will then be mistaken to be internal variability. In the three factor scaling method the underestimation is due to the opposite effect; when the phase of the internal variability lines up with the variability in MMM_{Nat} or MMM_{rest}, some of the internal variability will be removed. The two factor scaling method would appear to be the most accurate at estimating the standard deviation of the internal variability, although all the distributions are significantly different from the true distribution using a two-sided Kolmogorov–Smirnov test. This issue is explored further in Fig. 4.

In the detrending method the degree of overestimation correlates with the magnitude of the sensitivity to GHG (given by α_{GHG}), with large sensitivities leading to large estimates of natural variability (Fig. 4a). This is because large climate sensitivity results in highly nonlinear

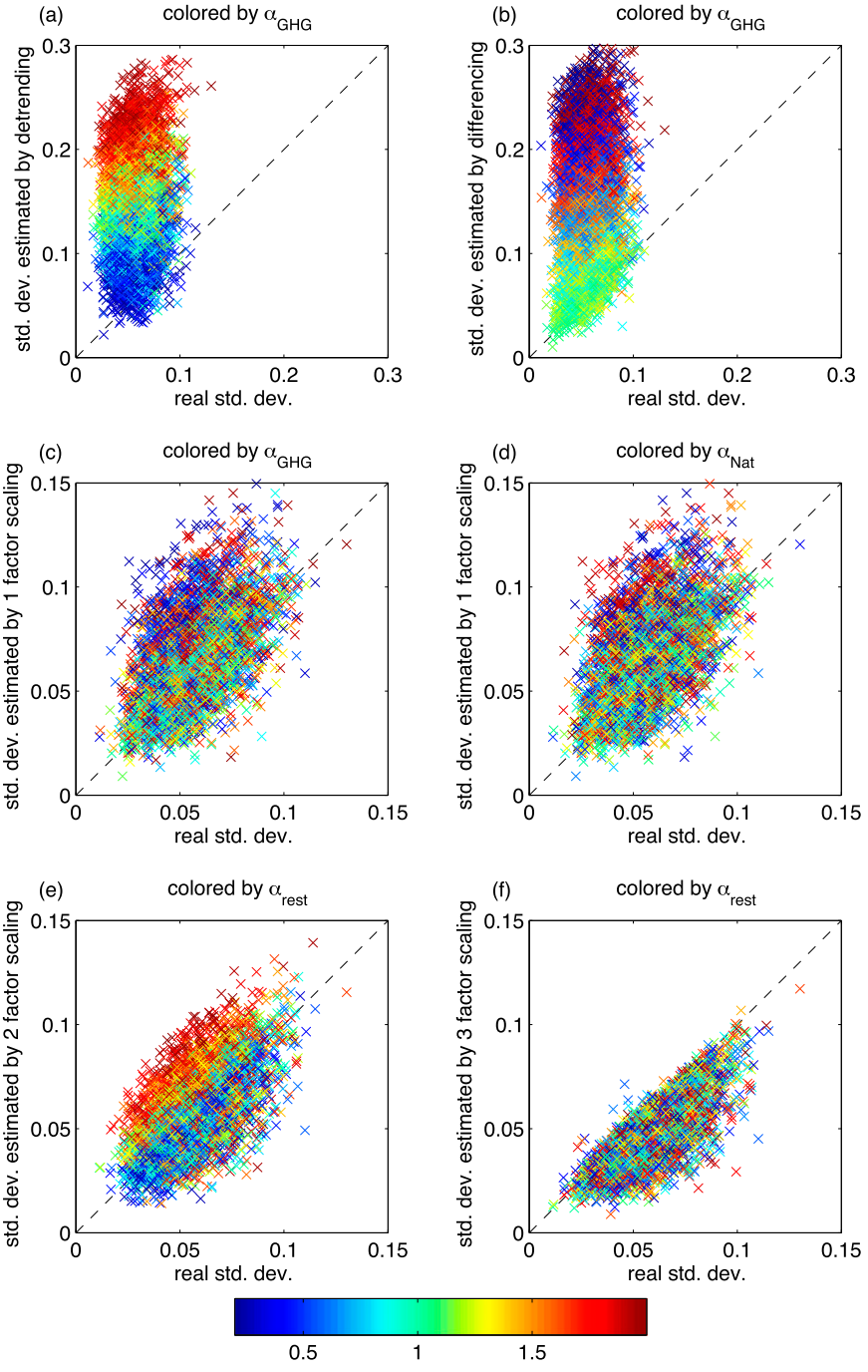

FIG. 4. Real vs estimated standard deviation of the synthetic time series for the (a) detrending, (b) differencing, (c),(d) single scaling, (e) two factor scaling, and (f) three factor scaling methods. In (a)–(c) color represents the (known) GHG scaling factor α_{GHG}, in (d) color represents the (known) natural scaling factor α_{Nat}, and in (e),(f) it represents the (known) residual scaling factor α_{rest}.

time series, for which a linear trend is a very poor approximation. In the differenced method it is the cases with either large or small values of α_{GHG} (dark blue and red crosses in Fig. 4b) that have the largest overestimation of amplitude because these are the cases for which the MMM is the poorest approximation of the forced signal. A similar, although less pronounced, bias occurs in the single scaling case (Figs. 4c,d); here it is the cases with a

large value of α_{GHG} and a small value of α_{Nat} (or conversely, a small value of α_{GHG} and a large value of α_{Nat}) that are overestimated. These are the cases for which the single scaling method will be the worst fit to the data because the single scaling method combines the sensitivity to GHG α_{GHG} and the sensitivity to natural forcings α_{Nat} in one parameter; it is thus a better approximation for cases where α_{GHG} and α_{Nat} are of similar magnitude. For two factor scaling (Fig. 4e) the amplitude in cases with large values of α_{rest} is overestimated, which is due to the misattribution of forced variability as internal variability as mentioned earlier.

Although it would appear from the distributions of standard deviations in Fig. 3d that the two factor scaling method may give a better estimate of the amplitude than the three factor scaling method, Fig. 4e shows that this apparent improvement is due to the fact that the two factor scaling method sometimes overestimates the real amplitude (because of neglecting α_{rest}, causing misattribution of the forced signal as internal variability) and sometimes underestimates the real amplitude (because of misattribution of the internal variability as the forced signal). In contrast, the three factor scaling method (Fig. 4f) gives a tighter estimate of the amplitudes, with a bias toward underestimation resulting from misattribution of internal variability as the forced signal.

In summary, detrending and differencing, which are the simplest and most commonly used methods of removing the forced signal, both give large biases in the estimated amplitude of the variability, with detrending also causing large biases in the estimated phase. Differencing gives a better estimate of the phase during the earlier part of the time series, when GHG forcing is less important, but biases increase as GHG forcing becomes dominant. The scaling methods give more accurate estimates of the amplitude, although with one scaling factor there is a small overestimation of the amplitude because of the inability of the method to account for different models having different sensitivities to natural forcing. The two factor scaling method appears to accurately estimate the amplitude, but there are errors in the estimated phase resulting from not removing the portion of the signal because of forcings other than GHG and natural forcing (e.g., aerosols and ozone). Including this missing forcing as a third scaling parameter improves the estimate of the phase but leads to an underestimation of the amplitude resulting from misattribution of the internal variability as naturally forced variability (since they occur on the same time scales).

We note that our results provide what are presumably generous estimates of the accuracy of the scaling methods since the forced time series were constructed with the same MMMs that were then used to estimate the scaling factors. When applying these methods to more complex data we must be aware that the MMMs themselves are only estimates of the underlying structure of the time series. The difference between the MMM calculated from the model ensemble and the true forced signal of each model will likely introduce additional errors.

5. Application to CMIP5 simulations

We now apply the five different methods to the CMIP5 simulation results. In this case we do not know the underlying internal variability; however, we can compare the results of the five methods to the CMIP5 control runs, where external forcing is constant. We also do not know the underlying shape of the model response to the external forcing; we estimate it by the MMM from the GHG and natural forcing runs (whereas in the synthetic cases it was the MMMs by construction). We are thus implicitly assuming that the timing and relative amplitudes of the model responses are constant across the models (which is not necessarily true—e.g., some models may have a larger response to one type of natural forcing than another).

The mean and standard deviation of the CMIP5 NASSTI are shown in Fig. 5a in gray, along with the mean and standard deviations of the AMO indices after the various methods to remove the forced signal have been applied. The results are very similar to the synthetic data. Figure 5b shows the distribution of turning points for the CMIP5 data, and once again the results correspond closely to the synthetic data. In the raw time series the maxima and minima line up with the maxima and minima of the MMM (shown by the black triangles on the x axis). This bias is not improved by detrending (dark blue). The differencing method and single scaling methods both result in a reasonably even distribution of turning points, apart from the edge effects of the filter. The two factor scaling method, however, shows preferences for maxima around 1880 and 1940 and for minima around 1920 and 1970. The first and last of these peaks may be partially influenced by edge effects, but in the middle of the time series there is still clearly some bias in the phase related to turning points of MMM_{rest} (magenta triangles). The three factor scaling method, which does attempt to take the residual external forcing into account, also shows a uniform distribution of turning points. In reality the distribution of turning points of the AMOI may be nonuniform as a result of excitation of the variability by external forcings (Otterå et al. 2010; Zanchettin et al. 2012; Iwi et al. 2012; Menary and Scaife 2014). However, we see no evidence for that here; the lack of a common response across models may simply be

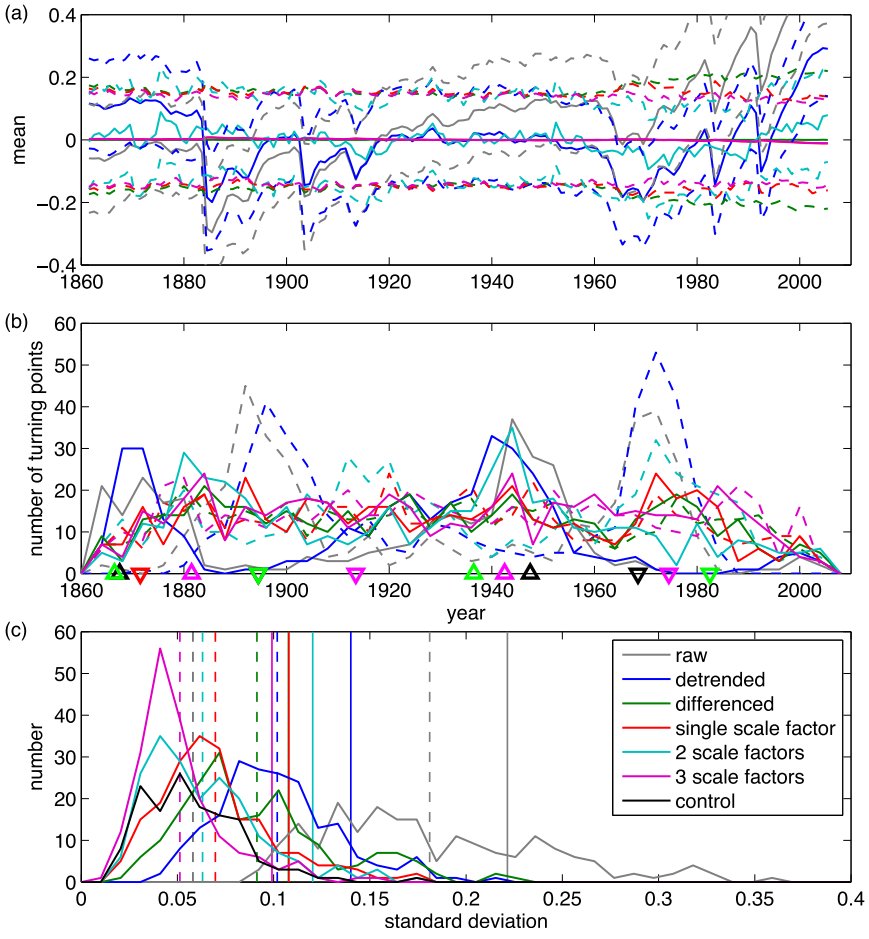

FIG. 5. (a) Time series of mean (solid) and one standard deviation either side of the mean (dashed) for the NASSTI time series (gray) and AMOI time series (colors). (b) Distribution of maxima (solid) and minima (dashed) as a function of time, with triangles on the *x* axis indicating minima and maxima of the MMMs as in Fig. 1. (c) Distribution of standard deviations of the amplitude of the NASSTI and AMOI. Dashed vertical lines indicate mean of the distributions while solid vertical lines indicate the standard deviation of the observed NASSTI (gray) and AMOI (colors). These may be compared to 145-yr-long sections of the control runs (black).

due to the different amplitudes, periods, and even mechanisms underlying North Atlantic climate variability in each model.

The distribution of amplitudes estimated by the various methods also follows the results found for the synthetic time series. In this case we also compare the amplitudes of internal variability estimated from the historical simulations to the amplitudes found in 145-yr-long sections of the control runs, where there is no variability in the external forcing (although note that control runs were not available for all models and that the amplitudes of variability from the control runs may be biased slightly high by slow drifts that can remain as a result of incomplete model spinup). The detrended, differenced, and, albeit to a lesser extent, single scale factor methods overestimate the amplitude of the internal variability while the three scale factor method underestimates it. The two scale factor method appears to give the best estimates of amplitude, as in the case of the synthetic data. Testing using a two-sided Kolmogorov–Smirnov test shows that the distribution of standard deviations from the control runs is not significantly different at the 99% level from the distributions calculated using the single scale factor and two scale factor methods.

Next we examine the scaling factors that are obtained from the regression of the CMIP5 NASSTI onto the various MMMs. These scaling factors indicate the sensitivity of each model to the various external forcings relative to the ensemble mean. For comparison we also calculate scaling factors for the observed NASSTI. Figure 6 shows the scaling factors for the single factor scaling method (Fig. 6a) and the three factor scaling

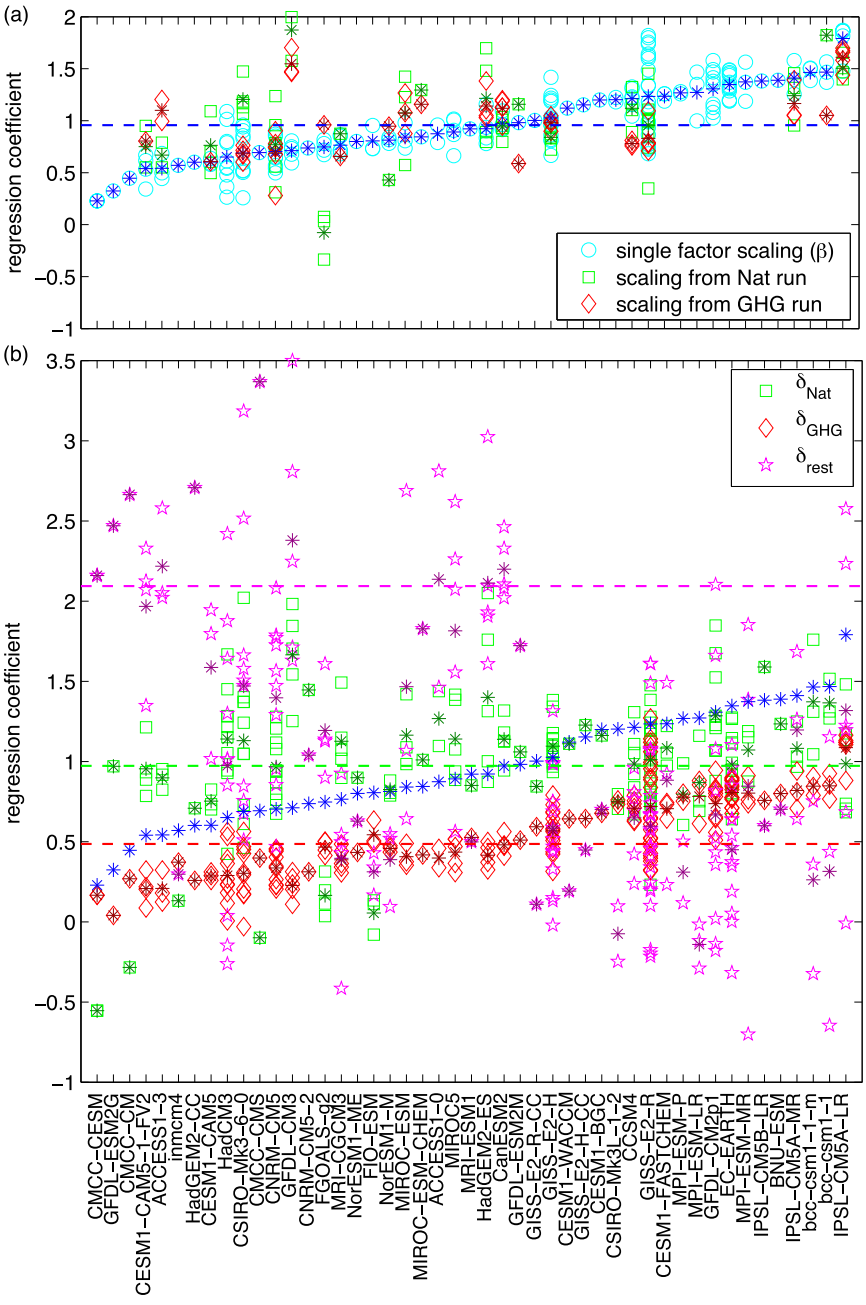

FIG. 6. (a) Regression (scaling) factors for the single factor scaling method for the all-forcing runs (blue) from CMIP5. Also included are the scaling factors obtained when scaling the natural forcing runs with MMM_{Nat} (green) and the GHG forcing runs with MMM_{GHG} (red), where those runs are available. (b) Scaling factors obtained using the three factor scaling method. Individual runs are plotted with shapes, and means for each model ensemble are shown with the asterisks. Horizontal dashed lines indicate the values obtained when applying the same scaling methods to the observed NASSTI. The blue asterisks from (a) are repeated in (b) for comparison.

method (Fig. 6b), along with the corresponding values for observations (dashed lines). We can see that there is a correlation between the scaling from the single factor scaling method and the GHG scaling factor from the three factor scaling method (red and blue asterisks in Fig. 6b), indicating that GHG sensitivity dominates the single factor scaling, as was the case with the synthetic data. Another estimate of the natural and GHG

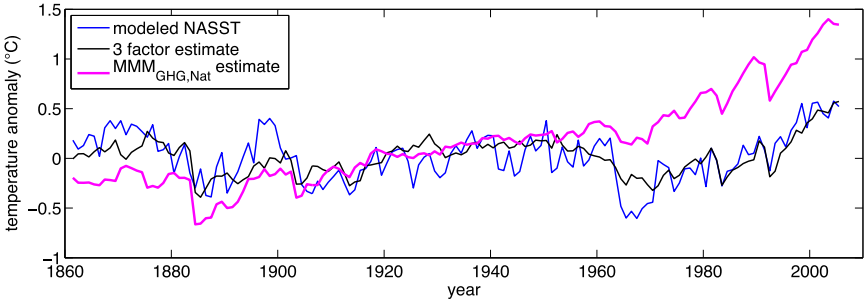

FIG. 7. Estimate of the forced signal from one ensemble member from the GFDL CM3 model, showing the impact of missing forcing factors. The modeled NASSTI is shown in blue. Also shown are two estimates of the forced signal, one using the three factor scaling method (black) and the other using scaling from the GHG-only and natural-forcing-only runs (magenta).

sensitivities can be made by regressing MMM_{Nat} and MMM_{GHG} on to each model's natural only and GHG-only forcing runs. These estimates are shown in Fig. 6a. There is, however, little or no correlation between the GHG scaling factor from the three factor scaling method and the GHG scaling factor obtained directly from the GHG-only forcing runs (cf. red asterisks in Figs. 6a,b; also Fig. 8a). Note that the GHG scaling factor from the three factor scaling method is less than unity for most ensemble members, indicating that the estimates of GHG sensitivity obtained from the all-forcing runs are generally lower than the estimates obtained from the GHG-only runs. This systematic difference is due to the all-forcing scenarios containing forcings, such as anthropogenic aerosols, that are not included in the GHG-only runs but that have time series with a significant projection onto MMM_{GHG} (Andreae et al. 2005). Anthropogenic aerosols act to partially offset GHG-induced warming, and thus the runs that include aerosol forcing will have a lower sensitivity since δ_{GHG} now represents sensitivity to GHGs combined with other forcings rather than the sensitivity to just GHGs alone.

As an illustration of the impact of the missing forcings on the estimates of the sensitivity parameters, Fig. 7 compares different estimates of the forced signal for one particular ensemble member from the GFDL CM3 model (Griffies et al. 2011). This model shows large sensitivities to both GHG and natural forcing when those sensitivities are estimated from the GHG-only and natural-forcing-only runs; however, an estimate for the forced time series made using those individual independent forcing sensitivities (magenta line in Fig. 7) is not a good fit for the modeled NASSTI (blue line). The three factor scaling method (in black) using the sensitivities from the all-forcing run provides a much closer fit using a lower estimate of the GHG sensitivity since that sensitivity is now no longer to GHG alone but includes other forcings that project significantly onto MMM_{GHG}. Other models show similar results (Fig. 8a).

The natural forcing scaling factor agrees better with the value obtained from the natural forcing runs (green asterisks in Figs. 6a,b; also Fig. 8b). There is a wider spread in the estimated values of the natural scaling factor compared to the estimates of the GHG scaling factor, with some models even having negative values (i.e., the opposite response to the forcing than the MMM). Part of this spread is due to the inaccuracy of the method since we know from the synthetic data that there can be larger errors in estimating the natural scaling factor than the GHG scaling factor (see Fig. 2). Similarly, the scaling factors for the forced variability unaccounted for by natural and GHG forcing (magenta stars in Fig. 6) show a wide range, with some models giving negative values.

There is no correlation between the models' estimated sensitivity to GHG and their estimated sensitivity to natural forcings (Fig. 9), which justifies the choice of independent scaling parameters for the synthetic time series in section 4. This lack of correlation also highlights the limitations of the single scaling method, which uses the same scaling factor to account for both GHG and naturally forced responses. The fact that the single scaling method still provides good estimates of both phase and amplitude is due to the dominance of the GHG forcing over the natural forcing.

6. Application to observations

We have also applied the five different methods to the observed NASSTI, from 1870 to 2005, as shown in Fig. 10 (with the scaling factors used shown in Fig. 6). The largest differences between the various AMOI estimates occur toward the end of the record, with a spread of 11 years in the estimated timing of the most recent minimum (1976 for the raw NASSTI time series, 1978

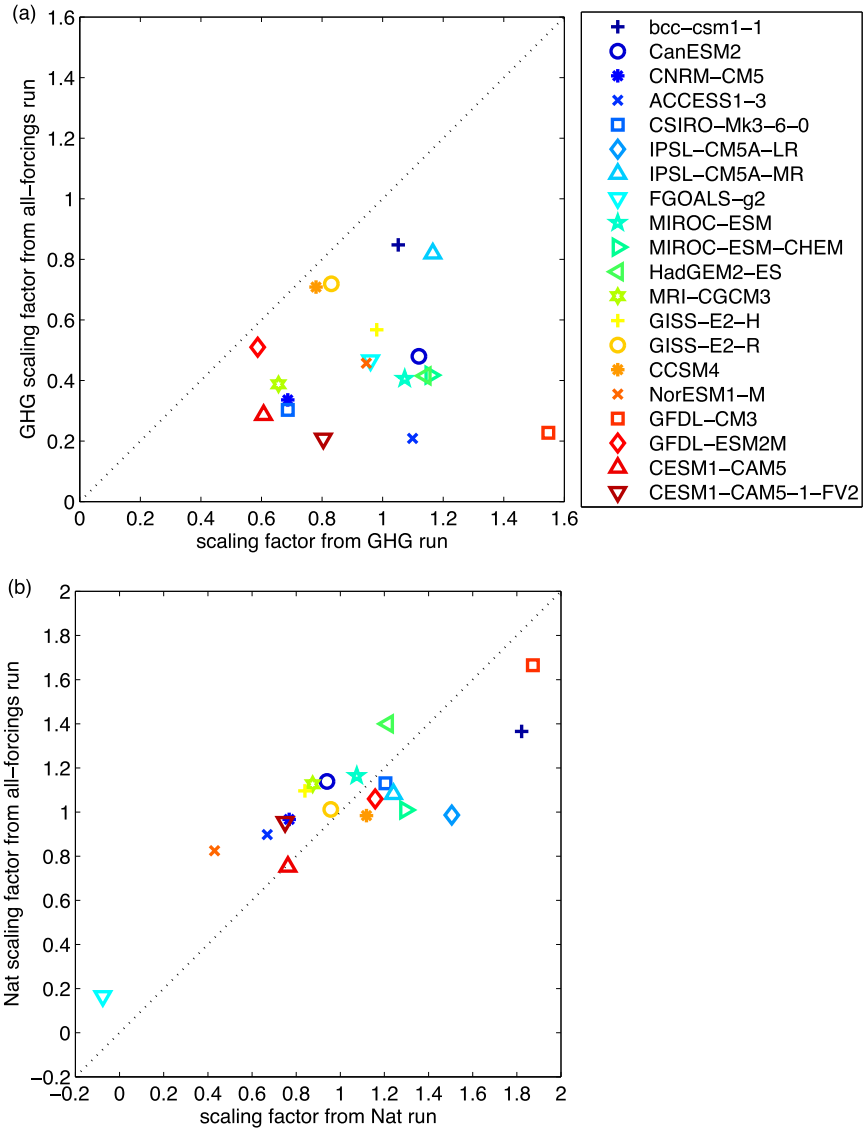

FIG. 8. Scatterplots of (a) scaling factors for GHG obtained from the GHG-only run compared to those obtained from the all-forcing run using the three factor scaling method and (b) scaling factors for natural forcings obtained from the natural-forcing-only run compared to those obtained from the all-forcing run using the three factor scaling method. The scaling factors are averaged over the ensemble for models where more than one ensemble member was available.

for the detrended time series, and 1987 for the three factor scaling time series, with the others in between). This in turn affects the estimated time of the predicted future maximum. The timing of the AMO (and other low-frequency modes of variability to which these methods may be applied) is important in ascertaining the role that the various modes of internal variability may be playing in the current and near-term future climate—for example, their relative contributions to the recent hiatus in the global-mean surface temperature increase.

Note that when applying the scaling methods to the observations we still use the MMMs from the models. Since the CMIP5 all-forcing, GHG, and natural forcing runs extend only until 2005 it is not possible to extend the time series in Fig. 10 without making further assumptions (e.g., persistence of the mean or trend; see Steinman et al. 2015). Also note that the phase is estimated using smoothed time series, so edge effects may be important. This means that estimates near the end of the time series may change as additional data become available.

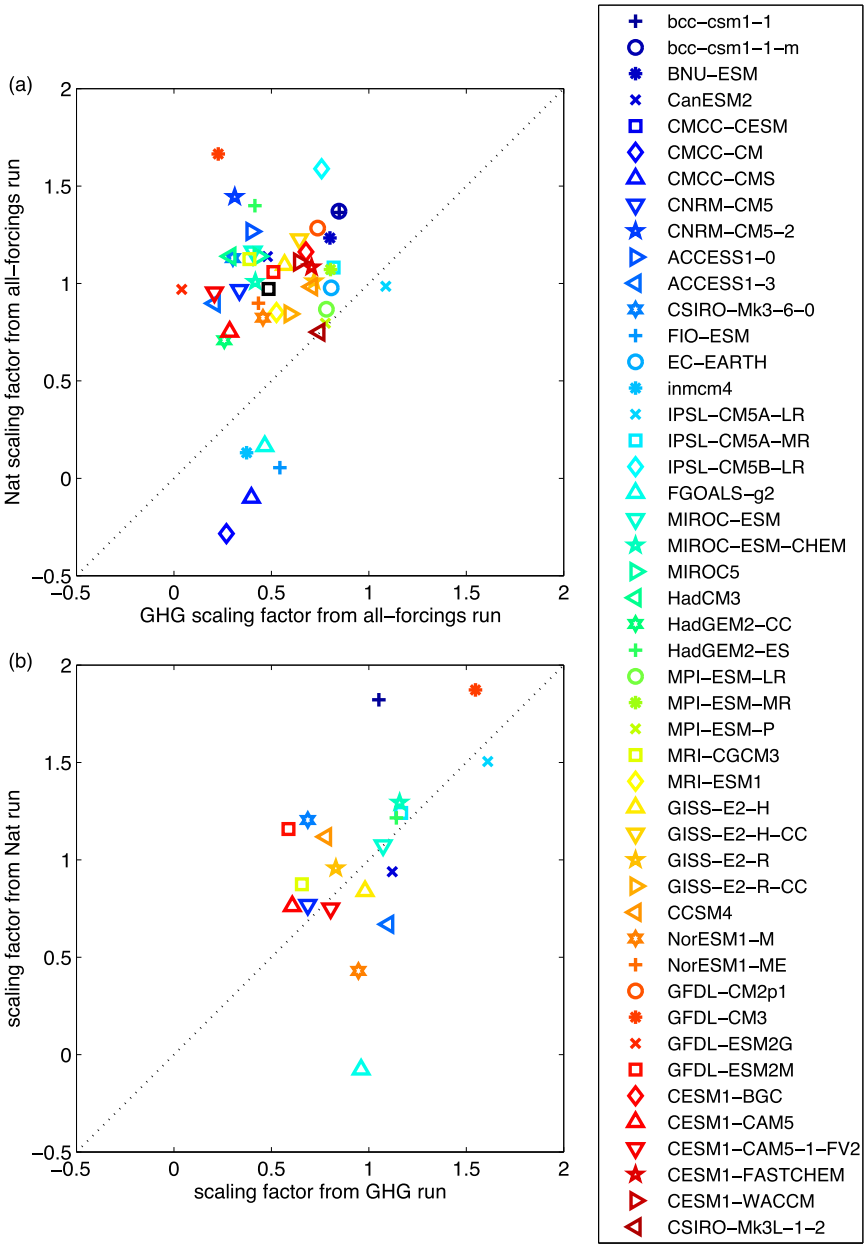

FIG. 9. Scatterplots of scaling factors (a) for GHG compared to scaling factors for natural forcings obtained from the all-forcing runs using the three factor scaling method and (b) from the GHG-only and natural-forcing-only runs (for those models where data is available). The scaling factors are averaged over the ensemble for models where more than one ensemble member was available. The scaling factors for observations are plotted in (a) as a black square.

Comparing the scaling factors obtained from observations (dashed lines in Fig. 6 and black square in Fig. 9) to those from the models, we can see that the observed GHG, natural, and residual scaling factors are all within the range simulated by the CMIP5 models.

Comparing the estimated amplitudes of the observations to the estimates from the models, we can see in Fig. 5 that for the three factor scaling method the observations have an amplitude greater than about 95% of the model ensemble members, suggesting that many of the models may not be simulating multidecadal variability of large enough amplitude in the North Atlantic. There is also the possibility of underestimation of the amplitude (in both the CMIP5 results and observations) using this method. It is already known, however, that models tend to underestimate decadal variability in the

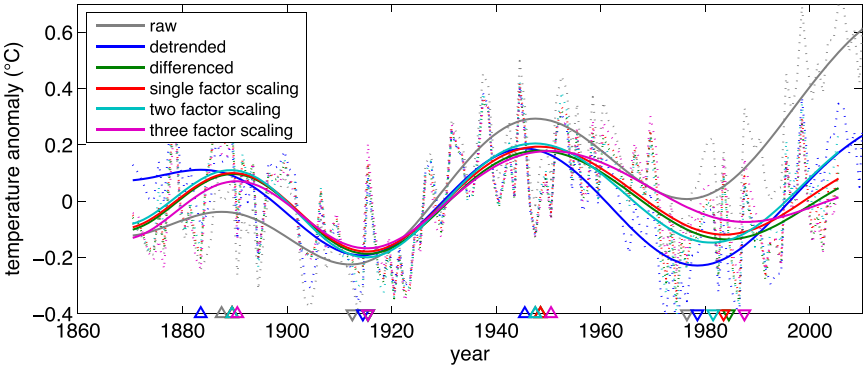

FIG. 10. Observed AMO indices calculated using the five different methods (colors) as well as the raw NASSTI (gray). Dotted (solid) lines indicate the raw (40-yr smoothed) time series. Upward (downward) pointing triangles along the *x* axis mark the position of maxima (minima) of the smoothed time series.

Pacific (e.g., England et al. 2014); perhaps this is a problem that also applies to decadal modes of variability in other ocean basins.

7. Main sources of error in the methods to estimate the forced signal

a. MMM shape

One major difference between the synthetic time series analyzed in section 4 and the CMIP5 models analyzed in section 5 is that for the synthetic time series the MMMs were known exactly because they were used in the construction of the time series. For the CMIP5 models, each different model will have a slightly different ensemble mean, and while the differences between these ensemble means and the MMM (constructed using all the models) are minimized using the scaling, they are not completely removed. Figure 11 shows ensemble means from the natural-forcing-only runs for five models (each of which has five or more ensemble members). Comparing the ensemble means to the MMM (in black) and taking into account the noise of each ensemble mean resulting from the smaller size of the ensembles compared to the multimodel mean, we can see that each of the models has a slightly different forced response. Part of this may be due to differing sensitivity of the models to different components of the natural forcing or to different timing of the response in different models. In addition there is also the fact that different models may include different forcings or even the same forcings but implemented in different ways. For example, models with interactive atmospheric chemistry may simulate a volcanic eruption by directly adding aerosols to their atmospheres, whereas another model with a simpler atmosphere might simulate the same eruption by varying incoming radiation. These model differences will have an effect on the model responses. In addition, there are fewer ensemble members available for the GHG and natural forcing runs than for the all-forcing runs, making $\mathrm{MMM_{GHG}}$ and $\mathrm{MMM_{Nat}}$ less robust estimates than $\mathrm{MMM_{all}}$.

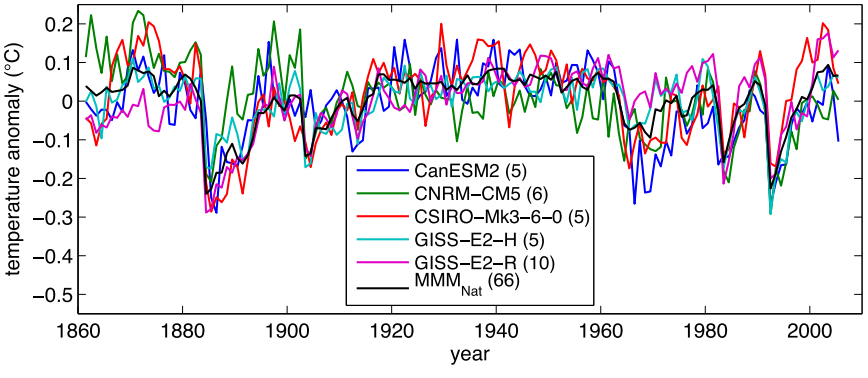

FIG. 11. MMM (black) and single model means (colors) for five different models forced with natural forcing only (chosen because each model had five or more ensemble members). The number of ensemble members included in each mean is shown in parentheses in the legend.

The same problems apply when using the MMMs to estimate the internal variability from observations (as in section 6), since we are assuming that the model MMMs adequately represent the true forced climate signal.

b. *Missing forcing factors*

Another factor worth considering more closely is the missing forcing types. Since we have GHG-only and natural-forcing-only runs available we have been able to attempt to account for these forcing types, but there are other forcings that may be important as well. Anthropogenic aerosols and ozone are of particular interest because they vary on time scales of the same order as the internal variability in which we are interested. Not taking a forcing into account leads to a spread in the estimated amplitude of the internal variability since, depending on the timing of the missing forcing, it may be either amplifying or canceling out the internal variability. This can be seen in Fig. 4e, where using only two scaling factors cannot account for the influence of the residual forcing given by α_{rest}. Including many different types of forcing leads to other problems, however, since time series may end up being overfitted, such that the true internal variability is mistaken as the forcing signal (as in Fig. 4f, where there is no problem with overestimation when α_{rest} is included but there is some underestimation).

Missing forcing factors are also responsible for the difference in estimating GHG scaling factors from the GHG-only runs and the all-forcing runs (which contain forcings that project onto the GHG forcing time series).

c. *Assumption of linearity*

Given enough computing power, both the above problems can be tackled by having more ensemble members and simulating more types and combinations of external forcings. However, a fundamental issue with all the methods described here is that we have assumed that the various forced signals and the internal variability can simply be combined linearly. Linearity was ensured by construction for the synthetic time series discussed in section 4. For the CMIP5 models this is not expected to introduce large errors since Schurer et al. (2013) found that the assumption of linearity held over the last millennium.

In addition, external forcing may have the ability to excite internal variability. However, we have not seen any evidence of this in our results (i.e., a bias toward a particular phase that cannot be explained by the limitations of the various methods).

d. *Possibilities for improvement*

As mentioned above, some of the challenges that arise in using the scaling method can be reduced using greater computing power. Having more ensemble members would provide more robust estimates of the various MMMs, and performing simulations for various forcings separately would allow more forcings to be included, although it would also increase the possibility of misattribution. Having more ensemble members for individual models would also allow individual model ensemble means to be used instead of MMMs, removing one potential source of error. Comparing to observations remains error prone, however, because of the necessary but imperfect assumption that the MMMs are applicable to the real world.

As for extending the methods into future projections, the different climate sensitivities mean that the different model trajectories diverge rather quickly as GHG forcing increases. Small errors in the estimated sensitivities at the end of the historical run quickly become overwhelming and make the estimates of internal variability in model forecasts increasingly unreliable. In addition, while the differencing and single scaling methods can be extended using RCP simulations, the two and three factor scaling methods rely on the Hist_{GHG} and Hist_{Nat} runs, which are available only until 2005.

8. Conclusions

The aim of this study was to assess the performance of methods for separating internal and forced variability of the climate, with application to North Atlantic sea surface temperatures. We have tried five methods: detrending, differencing, and three different scaling methods. Detrending, which is very commonly used in an attempt to remove the anthropogenic signal, leads to large overestimations of the amplitude of internal variability as well as large biases in the estimated phase of the variability, which can in turn bias the estimated period. Similarly, differencing (i.e., taking the difference between the observed climate and an estimate of the forced signal given by the multimodel mean from CMIP5) is not an ideal method. It gives a less biased estimate of the phase than simply detrending but still overestimates the amplitude of the variability because of different climate sensitivities of the different models.

Scaling the MMM responses to various types of forcing improves the estimates of the forced signal; however, care must be taken to include all the relevant forcings. Assuming that the models will have the same sensitivity to GHG and natural forcing (by using the MMM_{all} as in the single scaling method) improves the estimates of the phase and amplitude of the internal variability, although there can still be errors for models that have large sensitivity to GHG forcing and low sensitivity to natural forcing, or vice versa (Figs. 4c,d). The single scaling

method does, however, represent a significant improvement over the detrending or differencing methods. When GHG and natural forcings are scaled separately but the residual forcing is not included (as in the two factor scaling method) there can be either under- or overestimation of the amplitude of internal variability as well as a bias of the estimated phases toward the phase of the residual forced signal. Including the residual forcing (as in the three factor scaling method) improves the estimate of the phase but leads to a tendency toward underestimation of the amplitude. All the scaling methods suffer to varying extents from misattribution of the internal variability as the forced signal, which leads to underestimation of the amplitude when the phases of internal variability line up with the phases of the forced signal. The underestimation increases as more factors are included in the scaling. In addition, the scaling methods are subject to limitations, such as those due to the imperfect estimations of the various MMMs, variability due to missing forcings, and the assumption that the various forcings combine linearly. Despite these limitations, however, the scaling methods perform significantly better than detrending or differencing the time series. It is recommended that such scaling methods be used in preference to detrending or differencing in studies of low-frequency internal variability of the climate system.

Applying the five methods to observations suggests that many models may underestimate the amplitude of internal variability in the North Atlantic (with the caveat that the methods applied to both models and observations are prone to underestimation). The different methods lead to different results for the timing of the last minimum in the observed AMO index and thus different predictions for the recent/future maximum. These disparate predictions highlight the importance of being able to correctly distinguish between the externally forced signal and internal variability.

Acknowledgments. This work was supported by the Australian Research Council (ARC), including the ARC Centre of Excellence in Climate System Science. The authors acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and thank the climate modeling groups for producing and making available their model output. HadISST data were provided by the Met Office Hadley Centre (www.metoffice.gov.uk/hadobs).

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.

——, and P. A. Stott, 2003: Estimating signal amplitudes in optimal fingerprinting, part I: Theory. *Climate Dyn.*, **21**, 477–491, doi:10.1007/s00382-003-0313-9.

Andreae, M. O., C. D. Jones, and P. M. Cox, 2005: Strong present-day aerosol cooling implies a hot future. *Nature*, **435**, 1187–1190, doi:10.1038/nature03671.

Booth, B. B. B., N. J. Dunstone, P. R. Halloran, T. Andrews, and N. Bellouin, 2012: Aerosols implicated as a prime driver of twentieth-century North Atlantic climate variability. *Nature*, **484**, 228–232, doi:10.1038/nature10946.

Delworth, T. L., and M. E. Mann, 2000: Observed and simulated multidecadal variability in the Northern Hemisphere. *Climate Dyn.*, **16**, 661–676, doi:10.1007/s003820000075.

——, S. Manabe, and R. J. Stouffer, 1993: Interdecadal variations of the thermohaline circulation in a coupled ocean–atmosphere model. *J. Climate*, **6**, 1993–2011, doi:10.1175/1520-0442(1993)006<1993:IVOTTC>2.0.CO;2.

——, ——, and ——, 1997: Multidecadal climate variability in the Greenland Sea and surrounding regions: A coupled model simulation. *Geophys. Res. Lett.*, **24**, 257–260, doi:10.1029/96GL03927.

England, M. H., and Coauthors, 2014: Recent intensification of wind-driven circulation in the Pacific and the ongoing warming hiatus. *Nat. Climate Change*, **4**, 222–227, doi:10.1038/nclimate2106.

Flato, G., and Coauthors, 2013: Evaluation of climate models. *Climate Change 2013: The Physical Science Basis*, T. F. Stocker et al., Eds., Cambridge University Press, 741–866. [Available online at https://www.ipcc.ch/pdf/assessment-report/ar5/wg1/WG1AR5_Chapter09_FINAL.pdf.]

Folland, C. K., D. E. Parker, and F. E. Kates, 1984: Worldwide marine temperature fluctuations 1856–1981. *Nature*, **310**, 670–673, doi:10.1038/310670a0.

——, T. N. Palmer, and D. E. Parker, 1986: Sahel rainfall and worldwide sea temperatures, 1901–85. *Nature*, **320**, 602–607, doi:10.1038/320602a0.

Frankcombe, L. M., S. McGregor, and M. H. England, 2015: Robustness of the modes of Indo-Pacific sea level variability. *Climate Dyn.*, **45**, 1281–1298, doi:10.1007/s00382-014-2377-0.

Griffies, S. M., and Coauthors, 2011: The GFDL CM3 coupled climate model: Characteristics of the ocean and sea ice simulations. *J. Climate*, **24**, 3520–3544, doi:10.1175/2011JCLI3964.1.

Huber, M., U. Beyerle, and R. Knutti, 2014: Estimating climate sensitivity and future temperature in the presence of natural climate variability. *Geophys. Res. Lett.*, **41**, 2086–2092, doi:10.1002/2013GL058532.

Huck, T., A. Colin de Verdière, and A. J. Weaver, 1999: Interdecadal variability of the thermohaline circulation in box-ocean models forced by fixed surface fluxes. *J. Phys. Oceanogr.*, **29**, 865–892, doi:10.1175/1520-0485(1999)029<0865:IVOTTC>2.0.CO;2.

Iwi, A. M., L. Hermanson, K. Haines, and R. T. Sutton, 2012: Mechanisms linking volcanic aerosols to the Atlantic meridional overturning circulation. *J. Climate*, **25**, 3039–3051, doi:10.1175/2011JCLI4067.1.

Kerr, R. A., 2000: A North Atlantic climate pacemaker for the centuries. *Science*, **288**, 1984–1985, doi:10.1126/science.288.5473.1984.

Knight, J. R., R. J. Allan, C. K. Folland, M. Vellinga, and M. E. Mann, 2005: A signature of persistent natural thermohaline circulation cycles in observed climate. *Geophys. Res. Lett.*, **32**, L20708, doi:10.1029/2005GL024233.

Kushnir, Y., 1994: Interdecadal variations in North Atlantic sea surface temperature and associated atmospheric conditions. *J. Climate*, **7**, 141–157, doi:10.1175/1520-0442(1994)007<0141:IVINAS>2.0.CO;2.

Mann, M. E., 2008: Smoothing of climate time series revisited. *Geophys. Res. Lett.*, **35**, L16708, doi:10.1029/2008GL034716.

——, and J. Park, 1994: Global-scale modes of surface temperature variability on interannual to century timescales. *J. Geophys. Res.*, **99**, 25 819–25 833, doi:10.1029/94JD02396.

——, and K. A. Emanuel, 2006: Atlantic hurricane trends linked to climate change. *Eos, Trans. Amer. Geophys. Union*, **87**, 233–241, doi:10.1029/2006EO240001.

——, J. Park, and R. S. Bradley, 1995: Global interdecadal and century-scale climate oscillations during the past five centuries. *Nature*, **378**, 266–270, doi:10.1038/378266a0.

——, B. A. Steinman, and S. K. Miller, 2014: On forced temperature changes, internal variability, and the AMO. *Geophys. Res. Lett.*, **41**, 3211–3219, doi:10.1002/2014GL059233.

Menary, M., and A. Scaife, 2014: Naturally forced multidecadal variability of the Atlantic meridional overturning circulation. *Climate Dyn.*, **42**, 1347–1362, doi:10.1007/s00382-013-2028-x.

Otterå, O. H., M. Bentsen, H. Drange, and L. Suo, 2010: External forcing as a metronome for Atlantic multidecadal variability. *Nat. Geosci.*, **3**, 688–694, doi:10.1038/ngeo955.

Parker, D., C. Folland, A. Scaife, J. Knight, A. Colman, P. Baines, and B. Dong, 2007: Decadal to multidecadal variability and the climate change background. *J. Geophys. Res.*, **112**, D18115, doi:10.1029/2007JD008411.

Rayner, N. A., D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent, and A. Kaplan, 2003: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res.*, **108**, 4407, doi:10.1029/2002JD002670.

Schurer, A. P., G. C. Hegerl, M. E. Mann, S. F. B. Tett, and S. J. Phipps, 2013: Separating forced from chaotic climate variability of the past millennium. *J. Climate*, **26**, 6954–6973, doi:10.1175/JCLI-D-12-00826.1.

Steinman, B. A., M. E. Mann, and S. K. Miller, 2015: Atlantic and Pacific multidecadal oscillations and Northern Hemisphere temperatures. *Science*, **347**, 988–991, doi:10.1126/science.1257856.

Sutton, R. T., and D. L. R. Hodson, 2003: Influence of the ocean on North Atlantic climate variability 1871–1999. *J. Climate*, **16**, 3296–3313, doi:10.1175/1520-0442(2003)016<3296:IOTOON>2.0.CO;2.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bull. Amer. Meteor. Soc.*, **93**, 485–498, doi:10.1175/BAMS-D-11-00094.1.

Ting, M., Y. Kushnir, R. Seager, and C. Li, 2011: Robust features of Atlantic multi-decadal variability and its climate impacts. *Geophys. Res. Lett.*, **38**, L17705, doi:10.1029/2011GL048712.

Zanchettin, D., C. Timmreck, H. F. Graf, A. Rubino, S. Lorenz, K. Lohmann, K. Krüger, and J. Jungclaus, 2012: Bi-decadal variability excited in the coupled ocean–atmosphere system by strong tropical volcanic eruptions. *Climate Dyn.*, **39**, 419–444, doi:10.1007/s00382-011-1167-1.

Zhang, L., and C. Wang, 2013: Multidecadal North Atlantic sea surface temperature and Atlantic meridional overturning circulation variability in CMIP5 historical simulations. *J. Geophys. Res. Oceans*, **118**, 5772–5791, doi:10.1002/jgrc.20390.

Zhang, R., and Coauthors, 2013: Have aerosols caused the observed Atlantic multidecadal variability? *J. Atmos. Sci.*, **70**, 1135–1144, doi:10.1175/JAS-D-12-0331.1.