On semi-empirical decomposition of multidecadal climate variability into forced and internally generated components

Sergey Kravtsov* and David Callicutt

Department of Mathematical Sciences, University of Wisconsin-Milwaukee, WI, USA

ABSTRACT: This paper combines CMIP5 historical simulations and observations of surface temperature to investigate relative contributions of forced and internal climate variability to long-term climate trends. A suite of estimated forced signals based on surrogate multi-model ensembles mimicking the statistical characteristics of individual models is used to show that, in contrast to earlier claims, scaled versions of the multi-model ensemble mean cannot adequately characterize the full spectrum of CMIP5 forced responses, due to misrepresenting the model uncertainty. The same suite of multiple forced signals is also used to derive unbiased estimates of the model simulated internal variability in historical simulations and, after appropriate scaling to match the observed climate sensitivity, to estimate the internal component of climate variability in the observed temperatures. On average, climate models simulate the non-uniform warming of Northern Hemisphere mean surface temperature well, but are overly sensitive to forcing in the North Atlantic and North Pacific, where the simulations have to be scaled back to match observed trends. In contrast, the simulated internal variability is much weaker than observed. There is no evidence of coupling between the model simulated forced signals and internal variability, suggesting that their underlying dominant physical mechanisms are different. Analysis of regional contributions to the recent global warming hiatus points to the presence of a hemispheric mode of internal climate variability, rather than to internal processes local to the Pacific Ocean. Large discrepancies between present estimates of the simulated and observed multidecadal internal climate variability suggest that our ability to attribute and predict climate change using current generation of climate models is limited.

KEY WORDS AMO; climate oscillations; internal/forced variability

Received 10 November 2016; Revised 22 March 2017; Accepted 22 March 2017

1. Introduction

Analysis of scientific literature suggests that most researchers agree that a sizeable fraction of the observed 20th century climate warming may be due to human activity (Cook et al., 2013). It is less clear though whether there is such a consensus on the issue of exactly how much of the observed warming over the last few decades has been anthropogenic (see, e.g. Ghil and Vautard, 1991; DelSole et al., 2011; Wu et al., 2011, among others). The IPCC (2013) fifth assessment report states that: ‘It is extremely likely that human influence has been the dominant cause of the observed warming since the mid-20th century’. Mann et al. (2016) provided quantitative estimates of this (high) likelihood using a semi-empirical approach involving model simulations and observations of surface temperature (Frankcombe et al., 2015; Steinman et al., 2015a). Yet, such quantitative statements necessarily rely on how skillful the state-of-the-art climate-system models are in simulating the observed climatic variability; of particular importance is to assess the models’ potential to simulate internal low-frequency (multidecadal) climate variability that may arise in the climate system in the absence of changes in the external (anthropogenic and natural) forcing (Jolliffe and Stephenson, 2003; DelSole and Shukla, 2010). In this paper, we adopt a modified version of the Steinman et al.’s semi-empirical approach to estimate and compare the internally generated component of the observed multidecadal climate variability with the one simulated by the CMIP5 climate models (Taylor et al., 2012).

Separating the forced climate signal from internal climate variability using purely statistical methods generally relies on the assumption that these two types of variations possess their own distinct spatiotemporal signatures. These methods include the standard empirical orthogonal function analysis (EOF: Preisendorfer, 1988; Monahan et al., 2009), singular spectrum analysis (SSA; Ghil and Vautard, 1991; Elsner and Tsonis, 2006) and its multivariate extension M-SSA (Moron et al., 1998; Ghil et al., 2002; Jamison and Kravtsov, 2010; Wyatt et al., 2012; Kravtsov et al., 2014; Groth and Ghil, 2015; Groth et al., 2016), multi-taper spectral domain approach (Mann and Park, 1994,1999), empirical mode decomposition (Huang and Wu, 2008; Wu et al., 2011); discriminant analysis (Schneider and Held, 2001; DelSole and Tippett, 2007); optimal persistence analysis (DelSole 2001, 2006), among others. Comparison of the observed and
simulated space/time patterns detected by these methods serves to assess the models’ performance in simulating the observed climate signals and provides clues about dynamical sources of the observed climate variability.

In contrast to purely data-based signal processing techniques mentioned above, another class of detection and attribution methods makes a more extensive and immediate use of different fully coupled and partially coupled climate-model simulations designed to diagnose dynamical causes of the simulated variability (see, e.g. Zhang et al., 2007; Semenov et al., 2010). DeSole et al. (2011) derived the ‘internal multidecadal pattern’ (IMP) of climate variability from CMIP3 control runs by maximizing its average predictability time (APT; DeSole and Tippett, 2009). These authors also estimated the forced signal’s pattern as the leading discriminant that maximizes the ratio of variance in the forced simulations to that in the control runs of CMIP3 models (see also Ting et al., 2009). The two patterns derived from models were finally combined in a fingerprinting procedure (Hasselmann, 1997; Hegerl et al., 1997; Allen and Tett, 1999; Tett et al., 1999) to determine the forced and internal components of the observed climate variability, with the IMP component shown to contribute substantially to decadal climate trends and to exhibit decadal predictability (DeSole et al., 2013). Note that applying model-derived patterns to the observed data assumes, once again, that models adequately represent dynamics governing the observed climate variability, which, while certainly hoped for, is not guaranteed. In a sense, the consistency between the models and observations diagnosed in such detection and attribution studies might be a function of the procedural design, whereas possible model/data discrepancies (due to, e.g. possible absence of the real-world dynamical modes and their associated patterns in model simulations) may be effectively masked.

A more intuitive and easily interpretable way of addressing a mixture of the forced signals and internal climate variability in observations and model simulations – which does not depend on pattern recognition, and which we will adopt in the present study – is to exploit model ensembles. In particular, to get a naturally unbiased estimate of the forced signal (and the residual internal variability) in any given model, one is to run multiple simulations of this model under the identical history of external forcings for an ensemble of perturbed initial conditions. These simulations would share the same forced signal, but would have statistically independent, uncorrelated realizations of the internally generated variability. Single-model ensemble-mean time series (hereafter, SMEM) would thus be dominated by the forced signal, since different realizations of the internal variability would tend to cancel in taking the SMEM. The standard uncertainty of the forced signal estimation in this case is $\sigma/\sqrt{M}$, where $M$ is the number of simulations in the model ensemble and $\sigma$ is the standard deviation of the model’s internal variability. Since $M$ is typically not that large, one can also make use of the fact that the forcing time series and ensuing forced signals in the models are dominated by the response to slowly varying forcing and further reduce the uncertainty of estimated forced signal via smoothing of the individual models’ SMEMs. This would thus minimize the fingerprint of the interannual internal variability in the SMEM forced signal estimates. Note, however, that forced signals computed via smoothed SMEM with a small number of ensemble members $M$ would still not be identical to the true forced signals estimated with $M \to \infty$, and would contain, partly, the signal associated with unfiltered internal variability; consequently, the residual internal variability so estimated would itself have a lower variance compared to the true internal variability.

Steinman et al. (2015a) and Frankcombe et al. (2015), following Kravtsov and Spannagle (2008), Knight (2009) and Terray (2012), among others, used the CMIP5 multi-model ensemble mean (hereafter, MMEM) to estimate the simulated forced signal in regional climate indices. In general, using the MMEM for this purpose is a very good idea, since the total number of simulations in the multi-model ensemble ($M \sim 100$) is much larger than that in individual-model ensembles ($M \sim 5$), which reduces the MMEM-based forced signal uncertainty substantially as compared to SMEM under the assumption that all CMIP5 models represent equal samples of the same forced response. The caveat here is that in taking the MMEM, one also averages out, along with internal variability, the uncertainty associated with different forcing subsets and different physical parameterizations used in the models (hereafter, model uncertainty). Frankcombe et al. (2015) derived estimates of this uncertainty in synthetic data sets designed to mimic CMIP5 20th century runs, but stopped short of combining these uncertainty estimates with the actual estimated forced signals in CMIP5 simulations and observations. Meanwhile, Kravtsov et al. (2015) showed that the model uncertainty actually dominates the inferred ‘internal variability’ in the CMIP5 individual model simulations when the MMEM is used to define the forced signal. They further demonstrated that the forced signals based on the smoothed SMEMs provide much more accurate estimates (compared to MMEM) of the true forced signals and residual internal variability in individual model ensembles.

Figure 1 provides an illustration of these issues using a set of synthetic ‘model simulations’ with different forced signals and red-noise internal variability. In these examples, smoothed SMEM provides a faithful estimate of the true forced variability (and, hence, residual internal variability) for each case, whereas MMEM does well only for model 1, where the true forced signal (linear trend) happens to be close to the MMEM. For the other two ‘models’, the MMEM forced-signal estimate is heavily biased, and the resulting estimated ‘internal variability’ is dominated by the difference between these models’ true forced response and the MMEM.

The novel aspect of Steinman et al.’s attribution methodology compared to previous studies is in rescaling the raw MMEM signal using linear regression to best match the observations of a given climatic time series; hence, these authors termed their approach for estimating the
forced signal ‘semi-empirical’. Physically, the rescaling is meant to correct for biases in the models’ climate sensitivity. Steinman et al. (2015a) and Frankcombe et al. (2015) computed the internal component of observed climate variability by subtracting this rescaled MMEM signal from the raw observed time series of each climate index they considered. The resulting semi-empirical forced signal (and internal variability) estimates in Steinman et al. (2015a) were associated with narrow bootstrap-based error bars, which gave an impression of a striking consistency between the forced signals simulated by different CMIP5 models. If it were real, this consistency would also translate into high confidence of the corresponding estimates of the observed internal variability.

By contrast, we will show in the present paper that the above bootstrap-based estimates of MMEM signal uncertainty poorly characterize the actual forced-signal uncertainty, due to misrepresenting model uncertainty. We will argue that using a collection of forced signals — estimated here via smoothed SMEMs of individual models — is essential for proper representation of the forced-signal uncertainty in the CMIP5 multi-model ensemble. Note once again that, in contrast to the MMEM signal, the individual SMEM signals would be more substantially contaminated by unfiltered internal variability, due to small sample size of individual model ensembles. Using a perfect-model Monte-Carlo approach, we will quantify the magnitude of this contamination and demonstrate that it is in fact low compared to the model uncertainty. Finally, we will combine our estimates of the model uncertainty (essentially due to the difference between SMEMs of individual models) and SMEM averaging uncertainty (due to small size of individual model ensembles) to come up with a faithful representation of the forced signal and its total uncertainty in the CMIP5 multi-model ensemble. Subtracting these forced signals from the observed and model-generated data sets will provide the corresponding estimates of the internal variability. The main goal of this paper is to quantify the skill of CMIP5 ensemble as a whole, as well as that of individual models, in reproducing the magnitude and spatiotemporal characteristics of the internal component of observed climate variability.

Further presentation is organized as follows. Section 2 describes the input data and outlines our methodology. In Section 3, we discuss the structure and magnitude of the estimated forced-signal uncertainty associated with the MMEM and SMEM methods, and quantify these methods’ inherent biases. The SMEM-based bias corrected forced-signal estimation procedure is used in Section 4 to isolate forced signals and residual internal variability, along with requisite uncertainties, in observations and CMIP5 simulations, for three climate indices. We then compare the magnitudes of the observed and simulated internal variability over the entirety of the CMIP5 ensemble, as well as in individual models. Section 5 briefly summarizes our results and discusses their implications. Appendix S1 (Supporting information) contains technical details and further in-depth analyses of the main-text results, and addresses an implicit assumption of independence between the forced and internal variability, which underlies our methodology. It also presents the results based on sub-ensembles of CMIP5 models.

2. Data sets and analysis methodology

2.1. Data sets and CMIP5 sub-ensembles

We utilized the output from CMIP5 historical 20th century simulations for models with four or more ensemble members (Table 1) to analyse the simulated sea surface temperature (SST)-based climatic indices representing Atlantic Multidecadal Oscillation (AMO: Kerr, 2000; Enfield et al., 2001), Pacific Multidecadal Oscillation (PMO: Steinman et al., 2015a) and Northern Hemisphere mean surface temperature (NMO: Steinman et al., 2015a). The simulated SST indices, as well as their observed counterparts, were the same as used by Steinman et al. (2015a) and were downloaded from that manuscript’s supplementary website (www.meteo.psu.edu/holocene/
Table 1. CMIP5 20th century simulations with four or more realizations, resulting in 18 independent ensembles with the total of 116 simulations.

| Model # | Model acronym | Number of realizations | AMO (0.8) | PMO (0.57) | NMO (1.02) |
|---------|---------------|------------------------|-----------|------------|------------|
| 1.      | CCSM4         | 6                      | 0.68      | 0.47       | 0.78       |
| 2.      | CNRM-CM5      | 10                     | 1.15      | 0.76       | 1.22       |
| 3.      | CSIRO-Mk3-6-0p| 10                     | **1.10**  | **0.57**   | **1.21**   |
| 4.      | CanESM2       | 5                      | **0.81**  | 0.53       | 0.95       |
| 5.      | GFDL-CM2p1    | 10                     | 0.61      | 0.48       | 0.78       |
| 6.      | GFDL-CM3a     | 5                      | 0.80      | 0.21       | 0.91       |
| 7.      | GISS-E2-Hp1   | 6                      | 0.82      | 0.70       | 1.04       |
| 8.      | GISS-E2-Hp2   | 5                      | 1.03      | 0.72       | 1.21       |
| 9.      | GISS-E2-Hp3   | 6                      | 0.72      | 0.61       | 0.92       |
| 10.     | GISS-E2-Rp1   | 6                      | 0.8       | 0.70       | 1.11       |
| 11.     | GISS-E2-Rp2   | 6                      | 1.1       | 0.71       | 1.30       |
| 12.     | GISS-E2-Rp3   | 6                      | 0.48      | 0.64       | 0.93       |
| 13.     | GISS-E2-Rp4   | 6                      | 0.56      | 0.50       | 0.71       |
| 14.     | HadCM3        | 10                     | 0.83      | 0.57       | **1.11**   |
| 15.     | HadGEM2-ES3   | 5                      | **0.84**  | **0.33**   | **1.21**   |
| 16.     | IPSL-CM5A-LR  | 6                      | 0.48      | 0.42       | 0.72       |
| 17.     | MIROC5a       | 4                      | 0.93      | 0.64       | **1.13**   |
| 18.     | MRI-CGCM3a    | 4                      | 0.77      | 0.70       | 1.20       |

Models highlighted in italic/bold font are from the ‘poor’/‘good’ groups, respectively, per Martin et al. (2014). The last three columns list relative model sensitivities in terms of the scaling factors $b_{x\alpha}$ obtained via regression of SMEMs against the observed time series, for the AMO, PMO and NMO climate indices according to Equation (2). The numbers in parentheses show the multi-model average of the corresponding scaling factors. *Models have the aerosol indirect effects (cloud albedo+lifetime) included.

public_html/supplements/Science2015). The AMO and PMO indices were based on SST averaged over the regions (0°–60°N, 80°W–0°W) and (0°–60°N, 120°E–100°W), respectively. The NMO index was computed as the mean surface temperature (ocean+land) over the 0°–60°N region. The observed AMO and PMO indices were computed as the average of three SST products: the Hadley Centre Global Sea Ice and Sea Surface Temperature (HadISST: Rayner et al., 2003), National Oceanic and Atmospheric Administration (NOAA) Extended Reconstructed Sea Surface Temperature (ERSST) (Xue et al., 2003; Smith et al., 2008) and Kaplan SSTs (Parker et al., 1994; Reynolds and Smith, 1994; Kaplan et al., 1998). The NMO index was based on Goddard Institute for Space Studies (GISS) Surface Temperature (GISTEMP; Hansen et al., 2010; GISTEMP Team, 2016).

We also analysed separately three distinct sub-ensembles of CMIP5 simulations: the so-called ‘good’ and ‘poor’ multi-model sub-ensembles (highlighted by bold and italic fonts in Table 1) ranked, per Martin et al. (2014), by their ability to simulate the observed teleconnections between the Atlantic Multidecadal Variability (AMV) and Sahel rainfall, as well as the five available historical simulations of the HadGEM2-ES model (which is also a member of the ‘good’ group defined above) known to capture a large fraction of the AMV as the predominantly forced response (Booth et al., 2012). Based on these studies, the CMIP5 models differ in their estimates of the forced response, in part because they differ in their representation of complexity: note that most of the ‘good’ models (and none of the ‘poor’ models) have the aerosol indirect effects included in their physics package – as denoted by the asterisk in Table 1. These models also differ in the magnitude of their internally generated variability (see Evan et al., 2013; Martin et al., 2014 and references therein). In particular, some models are better at getting larger magnitude of internally generated variability because they include better representation of processes such as low stratocumulus cloud feedbacks (Evan et al., 2013) or dust feedbacks (Martin et al., 2014). Martin et al. (2014) also show that models with larger internally generated variability tend to be models with larger forced SST responses – presumably because the mechanisms that enable the models to capture more of the former are the same mechanisms/feedbacks that lead them to capture larger magnitude forced responses.

We examine these conjectures about possible dynamical dependencies between the CMIP5 simulated forced signal and internal variability in Section S6 of Appendix S1, using our estimates of the forced signal and internally generated variability in the entire CMIP5 ensemble and its targeted sub-ensembles, and find little evidence for any such relationship. These analyses thus provide a post-factum justification of our present treatment of the forced signal and internal variability as being completely independent.

2.2. Methodology

2.2.1. SMEM and MMEM methods

Consider multi-model ensemble of $M$ models, with each model having $K_m$ simulations ($m = 1, 2, \ldots, M$). Steinman et al. (2015a) compute an initial estimate of the forced signal, for a time series of given climate index $x$, as the MMEM $[x]$ of this index over the entirety of the
CMIP5 multi-model ensemble, and then rescale it using the least-square fit

\[ x_o = b_o \cdot [x] + a_o + \epsilon_o \]  

\[ f_o = b_o \cdot [x] + a_o \]  

With \( b_o \) representing the slope and \( a_o \) – the intercept of the estimated linear regression model, to ‘accommodate disparities in climate sensitivity’ between models and observations. In Equation (1), \( x_o \) is the observed time series of the climate index considered, \( f_o \) is the final semi-empirical estimate of the forced signal, and \( \epsilon_o \) is the final estimate of the internal component of observed variability. The uncertainty in \( f_o \) (and \( \epsilon_o \)) is estimated using bootstrap resampling of multi-model ensemble-member time series. We will show that the latter procedure misrepresents the model uncertainty.

To alleviate this problem, we employ an analogous methodology, but, instead of MMEM, use a smoothed SMEM of each individual model \([x_m]_1\):  

\[ x_o = b_{o,m} \cdot [x_m] + a_{o,m} + \epsilon_{o,m} \]  

\[ f_{o,m} = b_{o,m} \cdot [x_m] + a_{o,m} \]  

The smoothing is done with 5-year data-adaptive low-pass filter (Mann, 2008; available from www.meteo.psu.edu/holocene/public_html/smoothing08/). Kravtsov et al. (2015) showed that \([x_m]_1\) so computed defines an estimate of the forced signal in individual model ensembles that leads to uncorrelated samples of residual (simulated) internal variability. Our semi-empirical estimates of the forced-signal and internal variability components in the observed climatic time series \( x \) based on an individual model ensemble are thus \( f_{o,m} \) and \( \epsilon_{o,m} \), respectively.

Figure 2 illustrates the forced signal estimates defined above for the AMO, PMO and NMO indices. The raw (unscaled) 5-year low-pass filtered individual-model SMEM time series \([x_m]\) (multiple grey lines in the left panels) have a large spread characterizing the model uncertainty. Their ensemble mean \([x]\) for either AMO or PMO index exhibits a warm bias after the year 2000 with respect to the rescaled version of the ensemble mean \( f_\) [see Equation (1)]; the bias is slight (~0.1°C) in AMO and is more substantial (~0.3°C) in PMO. The simulated NMO signals exhibit essentially no bias. These biases or lack thereof are consistent with the scaling factors (relative sensitivities) \( b_{o,m} \) of individual models [see Equation (2)], which are listed in Table 1; the value of \( b_{o,m} \) less or greater than unity means that the corresponding model’s warming trend needs to be scaled down or up relative to observations. Such differences in the surface temperature response to forcing in different models may be due to a variety of factors; for example, due to model biases in the simulated climatological planetary boundary layer depth (Davy and Escau, 2016).

The grey lines in the right panels of Figure 2 correspond to the rescaled forced-signal estimates \( f_{o,m} \) based on individual model ensembles [see Equation (2)]. Note that the spread among these rescaled individual forced-signal estimates is naturally narrower than the spread among the non-scaled signals in the corresponding left panels, due to the individual estimates being nudged to the common observed time series. The remaining spread of \( f_{o,m} \) is still quite large. Note also that the forced signal estimate based on ensemble averaging the scaled smoothed SMEMs of individual models – i.e., taking \( f_{o,m} \) (thick coloured lines in the right panels of Figure 2) – is fairly close to the scaled MMEM \( f_o \) of raw (unscaled) smoothed SMEMs \([x_m]\) (thick black lines in the corresponding left panels of Figure 2); hence, interchanging the order of scaling and ensemble averaging does not drastically change the final estimate of the ‘most likely’ (ensemble-average) forced signal here.

2.2.2. SMEM method uncertainty

The uncertainty of the forced-signal estimate \( f_{o,m} \) based on each individual model ensemble can be computed by creating 100 synthetic Monte-Carlo versions of this ensemble. In particular, the time series \( x_{m,k} \) of the \( k \)th ensemble member of the \( m \)th model is decomposed as

\[ x_{m,k} = [x_m]_1 + \epsilon_{m,k} \]  

with \([x_m]\) – our smoothed SMEM – representing an estimate of this model’s forced signal, and \( \epsilon_{m,k} \) – its residual internal variability. Next, a low-order stochastic model forced by the Gaussian white noise is fitted and integrated to mimic the variance and lag-correlation structure of \( \epsilon_{m,k} \) (see Section S1 of Appendix S1). Synthetic versions of the internal variability \( \epsilon_{m,k} \) from the stochastic model are then combined with the original forced-signal estimate \([x_m]\) to produce the full synthetic ensemble time series of the individual model simulations

\[ x_{m,k}^{(s)} = [x_m] + \epsilon_{m,k}^{(s)} \]  

Finally, the 5-year low-pass filtered SMEM of \( x_{m,k}^{(s)} \) – \([x_m]^{(s)}\) is used in (2) instead of \([x_m]\) to define the corresponding semi-empirical estimates of \( f_{o,m} \) and \( \epsilon_{o,m} \). Repeating this procedure, say, 100 times, gives 100 estimated time series of \( f_{o,m} \) and \( \epsilon_{o,m} \); the ensemble means of these time series represent our final estimate of the forced signal and internal variability in observations based on the single model, and the spread of these times series about the ensemble mean characterizes the attributed uncertainty. Combining all 18 models considered in this way results in 1800 estimates of \( f_{o,m} \) and \( \epsilon_{o,m} \) and the grand multi-model ensemble-based semi-empirical estimates of the forced signal (and internal variability) in observations, as well as their uncertainty.

2.2.3. Comparing SMEM and MMEM methods

Monte-Carlo synthetic CMIP5 ensembles (Equation (4)) can also be used to compare the performance and elucidate inherent biases of the MMEM and SMEM forced-signal estimation methods in isolating the forced and internal variability in model simulations, since the ‘true’ forced signals \([x_m]\) and internal variability \( x_{m,k}^{(s)} \) in synthetic
samples are known in Equation (4) by construction. In the SMEM method, the estimated forced signal is the 5-year low-pass filtered SMEM $x_{m,k}$ that best matches the time series of a given simulation. In Steinman et al. method, the forced signal in an individual synthetic simulation $f_{m,k}$ is computed as in Equation (1), but using a given simulation’s time series $x_{m,k}^{(s)}$ in place of $x_{m,k}$:

\[
x_{m,k}^{(s)} = b_{m,k} \left[ x^{(s)} \right] + a_{m,k} + e_{m,k}^{(s)} \\
f_{m,k} = b_{m,k} \left[ x^{(s)} \right] + a_{m,k}
\]

Here $[x^{(s)}]$ is the synthetic MMEM, while coefficients $a_{m,k}$ and $b_{m,k}$ are computed to minimize the variance of $e_{m,k}^{(s)}$. The estimated forced signal $f_{m,k}$ is thus a scaled version of MMEM that best matches the time series of a given simulation. Note that in evaluating the performance of their MMEM method, Steinman et al. (2015a) defined the MMEM based on all multi-model simulations available, and Steinman et al. (2015b) only excluded the single simulation considered from the computation of MMEM forced-signal estimate for a given model simulation. In this paper’s application of MMEM method, we will exclude all of the simulations of a given ($m$th) individual model from the multi-model ensemble to compute this model’s unscaled forced signal $[x^{(s)}]$, and use the
ensemble-average over $K_m$ rescaled forced signals of this model in Equation (5b) – $[m_{x_k}]$ – to obtain the final single MMEM-based forced-signal estimate. This estimate should then be compared with the actual known forced signal $[x_k]$ in this model’s synthetic ensemble, or with the SMEM-based estimated version of this forced signal.

### 3. The uncertainties and biases associated with SMEM and MMEM methods to estimate forced and internal variability

For illustrative purposes, we consider here 100 synthetic multi-model AMO ensembles, with each ensemble consisting of 20 ‘models’ and 100 ‘simulations’ (five synthetic simulations per ‘model’); in particular, the estimated forced signals (5-year low-pass filtered SMEMs) from models 1 and 2 of Table 1 each enter the synthetic multi-model ensemble twice (since we only have 18 different forced signal estimates available from CMIP5 models).

Also, in this section, the single stochastic model fitted to the concatenated time series of all 116 original CMIP5 residuals (rather than separate stochastic formulations for 18 different CMIP5 models, as in Section S1 of Appendix S1) is used to produce synthetic internal variability realizations that are summed up with the SMEM-based forced signals to form full surrogate simulations. Section S6 of Appendix S1 utilizes analogous, but even larger synthetic ensembles of 200 ‘models’ and total of 1000 simulations per ‘multi-model’ ensemble to analyse the dependence of MMEM residuals on the number of simulations considered. All of these assumptions are relaxed in Section 4, where the configuration of each synthetic ensemble matches exactly that of Table 1 in terms of the number of models and simulations, and different stochastic models described in the appendix are used to mimic the properties of the internal variability of individual CMIP5 models.

#### 3.1. Quantifying errors of SMEM and MMEM methods

We first plot the variance of the raw and running-mean boxcar filtered time series of the actual and estimated internal variability based on SMEM and MMEM methods as a function of the filter’s window size (Figure 3(a)); the error bars show the 70% spread (approximately corresponding to ±1 standard deviation) of the variance among all of the synthetic simulations considered. Naturally, the variance decreases with the window size, as we apply more and more smoothing to the original raw time series. The SMEM-based internal residuals have variances that closely match the observed variances, and similar error bars; the estimated variance, however, is slightly lower than the observed variance due to aliasing of some of the actual internal variance into the forced signal. By contrast, the MMEM-based residuals have a much larger variance compared to the actual internal variance, and wide error bars. Note that the error bars of the SMEM and MMEM methods both overlap with known internal variability, but for the case of MMEM method this overlap merely indicates that for a few of the models, their true forced signal happens to be relatively close to the MMEM time series, whereas this is not the case for the majority of the models, as shown by a large ensemble-mean variance.

The differences between the two methods are better visualized in the error variance plots of the SMEM and MMEM-based estimates of internal variability (Figure 3(b)), which demonstrates that the SMEM method gives a much more accurate representation of the actual internal variability (much lower error variance) throughout the whole range of time scales considered. In relative terms, the SMEM method’s variance error is uniform, reflecting a reduction of less than 25% compared to the variance of the actual low-pass filtered internal variability for averaging window sizes exceeding 5 year (Figures 3(c) and (d)). On the other hand, the relative error of the MMEM regression-based estimates of internal variability becomes progressively larger with the averaging window size and already exceeds 100% for the 15-year low-pass filtered variability.

Since the estimated forced signal and internal variability add up to the same time series as the sum of actual forced and internal signals by construction, the errors in the internal variability illustrated in Figure 3 also characterize the errors in the forced-signal estimation. We end up with 100 estimates of the forced signal error time series for each of 20 models comprising the surrogate ensembles. We can further decompose the forced-signal errors into two parts. The model errors arise due to the inferred forced signal being systematically different from the actual forced signal of a given model across all of its 100 available realizations. To compute the model error, we averaged the forced-signal difference time series (inferred minus actual) over the 100 realizations, and then computed its root-mean-square (rms) average over time and across all 20 models. The remaining – averaging – errors are the rms complement to model errors. They arise due to insufficient cancellation/smoothing of internal variability in the ensemble-mean computation of the forced signal estimate.

The forced-signal estimation errors in the SMEM method are dominated by the averaging errors, and those in the MMEM method – by the model errors (Figure 4). Smaller averaging errors of the MMEM-based forced signal estimation are, of course, expected, due to a much larger number of independent realizations being averaged in forming the ensemble mean compared to the SMEM method. What is important though is that the MMEM method’s model errors drastically exceed the averaging errors of the SMEM method, which makes the latter the method of choice for estimating forced signals in the CMIP5 multi-model ensemble (Section 4).

The model errors in the MMEM regression estimates of the forced signal arise due to systematic differences between the true and estimated forced signal in each simulation. Indeed, Kravtsov et al. (2015) demonstrated that internal residuals within individual model ensembles exhibit high statistically significant correlations when the MMEM regression method is used to define the forced signal in these models. This fact is apparently at odds with Steinman et al.’s (2015a, 2015b) claim that the
Figure 3. Performance of the MMEM regression (red) and SMEM subtraction (blue) attribution methods in recovering the known internal variability (black) in 100 surrogate multi-model AMO data sets (see text). (a) Variance of actual and inferred internal signals; (b) error variance (variance of the difference between inferred and actual internal time series); (c) variance ratio (ratio of the black to the blue and black to the red lines in (a)); (d) relative error variance (error variance in (b) divided by the actual variance (black lines) in (a) times 100%). All of these characteristics were computed for raw and boxcar running-mean low-pass filtered time series using different window sizes of $2 \times K + 1$ year, $K = 0, 1, \ldots, 30$ (shown on the horizontal axis); $K = 0$ corresponds to raw annual data, $K = 1$ – to 3-year low-pass filtered data and so on. Error bars, where present, show the 70% spread (between 15th and 85th percentiles) of the quantity displayed across all of the surrogate simulations considered.

MMEM-based internal residuals are statistically independent. They reached this conclusion by forming the grand ensemble-mean time series of internal residuals and analysing its variance, with small variance supposedly indicating statistical independence. In reality, these arguments are flawed, and small variance of the ensemble-mean residuals instead reflects an algebraic constraint rooted in the definition of the MMEM forced signal (Kravtsov et al., 2015; Appendix S1, Section S7).

3.2. Uncertainties of forced-signal estimates based on multi-model ensemble

We can now make use of the surrogate forced-signal estimates computed using SMEM and MMEM methods to construct the most likely forced signal and compute its uncertainty (Figure 5). This most likely forced signal – ensemble mean of the unscaled SMEM-based forced signals (Figure 5(a), red curve) – is nearly identical to the MMEM (Figure 5(b), red curve). The spread of the individual SMEM forced signals (Figure 5(a), grey curves) reflects both model and averaging uncertainties, and is characterized by the standard deviation of about 0.1 °C. By contrast, the bootstrap-based spread of the MMEM estimates (Figure 5(b), grey curves) – used by Steinman et al. (2015a) to measure the uncertainty of the MMEM regression-inferred forced signal and internal variability – is much smaller (0.015 °C). The smallness of the MMEM uncertainty does not indicate that the MMEM provides a more robust estimate of the forced signal. Instead, it simply is a consequence of the fact that surrogate bootstrap ensembles, which typically contain about 2/3 of different model simulations from the all-model ensemble, effectively sample the forced signals from all of the models considered, thus resulting in a similar estimate of the
is larger (with the time-mean value of 0.093
mates resulting from the MMEM procedure (Figure 5(d))
tributed in time. The uncertainty in the forced signal esti-
The remaining forced-signal uncertainty is uniformly dis-
in Figure 5(a) due to eliminating the uncertainty associated
over the 100 multi-model ensembles considered, and then computing
the root-mean-square error for the resulting multi-model difference
time series. ‘Averaging’ errors (+-symbols) arise due to leaking a fraction
of the internal variance to the forced signal estimate due to insufficient
cancellation of independent realizations of internal variability in the
MMEM or SMEM ensemble means. The total attribution error (solid
lines) is the root-mean-square sum of the model and averaging errors.

MMEM every time. Hence, the narrow ranges of uncer-
tainty in semi-empirical estimates of the forced signal and
internal variability in Steinman et al. (2015a) are misleading,
as they ignore the model uncertainty.

Figures 5(c) and (d) show spaghetti plots of forced sig-
als obtained by combining the 5-year low-pass filtered
SMEM from individual models rescaled to best match the
MMEM, with the forced-signal error time series resulting
from the SMEM (Figure 5(c)) and MMEM (Figure 5(d))
methods. The forced signals from the SMEM procedure (Figure 5(c))
have a narrower spread (0.076°C) than that in Figure 5(a) due to eliminating the uncertainty associated
with the individual models’ different climate sensitivities.
The remaining forced-signal uncertainty is uniformly dis-
tributed in time. The uncertainty in the forced signal esti-
mates resulting from the MMEM procedure (Figure 5(d))
is larger (with the time-mean value of 0.093°C, that is,
with about 50% increase in the error variance compared to
that in Figure 5(c)) and has time-dependent structure elu-
cidating large forced-response biases in individual mod-
els. These results, once again, indicate that, despite small
size of individual-model ensembles, the SMEM method of
the forced signal inference provides narrower, more robust
and more structurally stable uncertainty estimates of the
forced signal and residual internal variability in individ-
cual CMIP5 models compared to the MMEM regression
method of Steinman et al. (2015a, 2015b) and Frankcombe
et al. (2015).

4. Observed and CMIP5 simulated internal
variability

In this section, we apply the SMEM Monte-Carlo method
of Section 2.2 to the CMIP5 multi-model ensemble
of Table 1. The model simulations span the period
1861–2005, and the observed time series extend through
2012. Following Steinman et al. (2015a), we linearly
extrapolated the synthetic forced-signal estimates result-
ning from our procedure through year 2012 using the
slope of 15-year (1991–2005) trend. Steinman et al.
used the 30-year trend for the same purpose, but the
(cors-validated retroactive hindcast) skills of the 15-year
and 30-year trend extrapolation are similar (not shown).
We computed two versions of the forced signals and their
uncertainties. In the first version (Figures 6(a), (c) and (e)),
we used the non-scaled forced signals. The uncertainty of
these estimates includes that due to different climate sen-
sitivities of the individual models. In the second version
(Figures 5(b), (d) and (f)), we rescaled the 1861–2005
surrogate forced signals according to Equation (2), and
then linearly extrapolated them through year 2012. We
reduced the 95% spread of the synthetic forced signals
dashed lines in Figure 6) by 0.032°C for AMO, 0.028°C
for PMO, and by 0.027°C for NMO; these reduction
factors were computed by comparing the 95% spreads of
the actual (known) and estimated internal variability in
the corresponding surrogate ensembles (4).

The spread of the non-scaled forced-signal estimates is,
naturally, the largest due to including the model-sensitivity

© 2017 The Authors. International Journal of Climatology published by John Wiley & Sons Ltd
Int. J. Climatol. 37: 4417–4433 (2017) on behalf of the Royal Meteorological Society.
4.1. Semi-empirical estimates of observed internal variability

Differencing the observed time series (purple lines in Figure 6) and our surrogate forced-signal estimates (grey lines in Figure 6) produces the corresponding surrogate estimates of the semi-empirical internal variability. Following Steinman et al. (2015a), we concentrated on its multidecadal component by applying to these time series the 40-year data-adaptive low-pass filter of Mann (2008) modified to alleviate its end effects (see Section S3 of Appendix S1). The resulting estimates of the tapered 40-year low-pass filtered observed internal variability are shown in Figure 7 for the entire multi-model ensemble and for the subset of surrogate simulations based on the HadGEM2-ES model. The all-model ensemble-mean estimates of the multidecadal internal variability in AMO, PMO and NMO in Figure 7 (left) are relatively similar to those in Steinman et al. (2015a) (see their figure 3(c)); yet, there are noticeable differences in the phase and magnitude of these variations. For example, our estimated AMO signal is decreasing during 2000–2012 period,
Figure 7. Estimates of observed multidecadal internal variability for AMO (a, b), PMO (c, d) and NMO (e, f). The estimates in the left panels were obtained using the Monte-Carlo ensembles of rescaled forced signals based on all models (as in Figure 6, right column). The estimates in the right panels were based on the synthetic ensembles resulting from the HadGEM2-ES model only. Rescaled forced signals were subtracted from the corresponding observed time series, 40-year low-pass filtered and windowed using the tapers shown in Figure S5 of Appendix S1. Heavy solid coloured lines (AMO: blue, PMO: green, and NMO: red) show the ensemble mean of the resulting internal signal estimates, and error bars – their 95% spread. Each panel also contains for reference the ‘internal’ estimates based on subtracting linear trend from the entire observed timeseries, as well as the one based on the piecewise linear detrending of the observed time series with the break point at 1900.

whereas Steinman et al.’s internal AMO estimate exhibits an increase during this period. Furthermore, our estimated NMO decrease during the same period is about twice as large as the Steinman et al. (2015a) estimate. Most importantly, though, our estimates of uncertainty are drastically different. Steinman et al.’s (2015a) narrow error bars are merely an artefact of bootstrap samples generation (see Section 3.2 and Figure 5), whereas the error bars in Figure 7 reflect the actual (large) uncertainty associated with both the model errors and averaging errors of the forced-signal estimation. These errors are sufficiently large to render the attribution of the recent cool down of the PMO (Figure 7(c)) and NMO (Figure 7(e)) to the internal variability barely statistically significant. In other
words, the uncertainty of the forced response in the CMIP5 ensemble is comparable with the magnitude of multi-decadal variations in the internal component of observed climate variability.

The above uncertainty is the product of differences in the shape and magnitude of estimated forced signals of the individual models. Consider, for example, HadGEM2-ES model (Figure 7, right panels). The HadGEM2-ES-based estimates of the observed internal variability and its associated single-model uncertainty are naturally within the multi-model spread shown in the corresponding left panels, for all of the three indices considered, but with a drastically different shape of the resulting multidecadal signal. In particular, this model suggests that the internal variability in the observed AMO index nearly disappeared in the second half of the 20th century (Figure 7(b)) – with the forced signal dominating the full observed variability over that period (Figure S3(b)), consistent with Booth et al. (2012). On the other hand, the HadGEM2-ES model’s estimate of the internal component of the observed PMO index is dominated by a non-uniform long-term trend. This is due to the fact that the non-scaled forced signal for this model has a shape very different from that of the observed PMO variability (Figure S3(c)). As a result, the rescaled forced signal estimate (Equation (2)), which minimizes the rms distance between the estimated forced time series and observations, is well approximated by a semi-horizontal line (Figure S3(d)), as reflected in the small value of the PMO scaling parameter $b_m = 0.33$ for this model (Table 1). Consequently, the semi-empirical internal variability estimated using HadGEM2-ES dominates the observed PMO time series. The HadGEM2-ES-based estimate of the internal component of the observed NMO is also very different from the result based on the entire CMIP5 ensemble (compare Figures 7(e) and (f)), due to large phase differences between the observed NMO time series and the HadGEM2-ES estimate of the forced signal (Figure S3(f)). The forced signal estimation results based on the ‘good’ and ‘poor’ sub-ensembles of models are in general more consistent between each other, as well as with the results based on all models (see Figures S1, S2 and S4 in Section S2 of Appendix S1), the largest differences between the two groups of models arising in the estimate of the NMO forced signal and the resulting semi-empirical estimate of the internally generated component of the observed NMO (Figures S4(c) and S4(f)).

An interesting message from Figures 7 and S4 relates to causes of the recent global warming hiatus – namely the slowdown of the observed warming in NMO time series over 1996–2012 period, presumably due to cancellation of the forced warming trend by the internally generated variability. Steinman et al. (2015a) argued that the hemispheric NMO hiatus was dominated by the downswing of the internal PMO variability. Indeed, their estimate of the internal PMO cooling between year 1996 and 2012 of about 0.3 °C is about twice as large as their estimated NMO cooling over that period, while the estimated internal AMO exhibits a warming trend (Steinman et al., 2015a; figure 3(c) there). A moderate NMO cooling could then be thought of as a diluted reflection of a strong PMO cooling signal in the Northern Hemispheric mean surface temperature. By contrast, our results show that the internal components of both of these indices have a comparable magnitude (1996–2012 drop of about 0.3 °C) if all-model ensemble-mean forcing is used to define the observed internal variability (Figures 7(c) and (e)). Moreover, in the sub-ensembles of models, the NMO recent drop actually exceeds that of PMO. For example, the ‘internal’ NMO drop of 0.5 °C estimated using HadGEM2-ES model (Figure 7(f)) is much larger than its estimated 0.2 °C PMO drop (Figure 7(d)), which makes it difficult to argue in this case that the slowdown of warming was caused by the Pacific temperature trend. The same kind of message arises from the analysis that uses ‘good’ model subgroup to define the forced signal (Figures S4(c) and (e)), whereas the estimated internal PMO and NMO drops from the sub-ensemble of ‘poor’ models are comparable, as in all-model results. It thus appears that the differences between the simulated and observed temperature trends over the past couple of decades enter at a hemispheric scale, rather than locally over the Pacific. This point is further rectified in Section S5 of Appendix S1 (Figure S7).

4.2. Magnitude of observed vs. simulated internal variability

We now have estimates of both observed (Section 4.1) and CMIP5 simulated internal variability, the latter obtained by subtracting the 5-year low-pass filtered SMEMs from individual model simulations. In the latter estimates, we need to correct for the amplitude bias due to forced-signal smoothing (see Figure 2(a)). We do so by utilizing our surrogate ensembles with known forced signals and internal variability and derive the frequency-dependent amplitude-correction factors by comparing the magnitudes of the actual and inferred internal variability in surrogate simulations. These factors amount to about 5% for raw annual data and saturate at about 8% for the low-pass filtered data (Figure 8(a)). They are applied to inflate the variance of the estimated internal variability derived from historical CMIP5 simulations and result in the estimates of the magnitude of internal variability consistent with those from the available pre-industrial control runs (black error bars); these runs are listed in Table S1.

Note that we only had the data available for 10 pre-industrial control runs. These 10 models, however, include most of the models with the internal variances in the upper echelon of the CMIP6 ensemble (Figure S9), so our estimates of the control-run internal variability based on the sub-ensemble of CMIP5 models are also likely to represent well the control-run variability of the entire CMIP5 ensemble. To obtain the final estimates of this variability, we pulled 1000 random 145-year-long segments from the 10 available control runs, and computed their standard deviations, for raw and low-pass filtered data. Black error-bar plots of Figure 8 show the (1000-member) ensemble-mean and standard uncertainty of the control-run STDs so computed.
Figure 8. Standard deviations (STDs) of the observed (blue) and CMIP5 simulated (red) internal variability in the AMO (b), PMO (c) and NMO (d) indices; the STDs spectra were computed for raw and boxcar low-pass filtered data as a function of the filter’s size, the latter shown on the abscissa of each panel (same conventions as in Figure 5). The input data for the semi-empirical estimates of the observed internally generated variability were the same as in the left column of Figure 7. The internally generated variability in the simulated time series was estimated by subtracting the 5-year low-pass filtered SMEM from individual model simulations. To account for the leakage of the ‘internal’ variance associated with the SMEM subtraction procedure, we computed the STD inflation factors as the ratio of the actual to the SMEM-inferred standard deviations of the (known) internal variability in the Monte-Carlo simulations of each climate index. The filter-size dependent inflation factors shown in panel (a) multiplied the standard deviations of the 116 raw CMIP5 internal signals; blue lines in (b–d) show these inflated standard deviations. The heavy lines in (b–d) show the ensemble-mean, and error bars – the associated 95% spread of the individual STDs. Black error bar plots show the mean and standard spread of the AMO, PMO and NMO’s standard deviations computed from the CMIP5 pre-industrial control runs (see text for details).

The upward corrected CMIP5 simulated internal variability in the AMO, PMO and NMO indices estimated from historical runs, as well as from pre-industrial control runs (Figures 8(c) and (d), respectively), all have standard deviations that are substantially smaller than the standard deviations of the observed internal variability (as also noted, for AMO, by Frankcombe et al., 2015). In particular, the ensemble-mean amplitudes of our estimates of the internal variability in observations exceed the 97.5th percentile of the simulated (inflated) amplitudes for a wide range of time scales, including the multidecadal variability. This means that internal variability simulated by the CMIP5 models is significantly weaker than the observed internal variability inferred by subtracting the CMIP5 derived scaled forced signals from the full observed climatic time series. Note that the internal component of the observed climate variability so defined has in fact minimum possible amplitude, as the forced signal rescaling minimizes, in the least-square sense, the deviations of the full observed time series from the model-derived forced signal; for example, the internal variability formed by differencing the observed time series with non-scaled forced signals, would have an even larger amplitude.

These results on magnitude of the internal variability based on the sub-ensembles of models (Section S4 and Figure S6) are generally consistent with those based on all models in that the internal variability in model simulations...
is significantly weaker than the semi-empirically derived internal component of the observed variability.

5. Summary and discussion

We considered three SST indices representing Northern Hemisphere climate variability (AMO, PMO, NMO) and developed a Monte Carlo approach to estimate uncertainties of their forced and internal components using the multi-model ensemble of CMIP5 20th century simulations. In particular, we generated, over the entirety of the CMIP5 multi-model ensemble considered, surrogate samples of individual model simulations by combining a given model’s forced signal estimated via smoothed single-model ensemble-mean (SMEM) with stochastic realizations mimicking the properties of each model’s internal variability.

We used these synthetic ensembles to show that, in contrast to claims by Steinman et al. (2015a, 2015b), it is impossible to characterize the entirety of the CMIP5 ensemble’s forced signals with the single multi-model ensemble-mean (MME) time series, since different models comprising this ensemble have drastically different shapes of their forced responses. Taking the MME averages over these different forced signals and results in drastic underestimation of the forced-signal uncertainty. Using, on the other hand, multiple SMEM-based forced signals allows one to obtain a much more truthful representation of the forced-signal uncertainty.

Ultimately, we used our synthetic ensembles to obtain unbiased estimates of the simulated internal variability based on historical CMIP5 runs, as well as semi-empirical estimates of internal component of the observed internal variability (the latter in conjunction with rescaling of the SMEM-based forced-signal estimates to match the observed climate sensitivity; see Steinman et al. 2015a; Frankcombe et al. 2015). The levels of internal variability in CMIP5 models do not exhibit any apparent association with the forced signal, which provides a post-factum justification for our treating the forced signals and internal variability as independent. Observed internal variability is characterized by a pronounced multidecadal signal of a hemispheric scale, which may have contributed to the 1996–2012 global warming ‘hiatus’. On the other hand – and this is, perhaps, the most important result of this paper – the CMIP5 simulated multidecadal variability is weaker than the inferred internal component of the observed variability by a factor of 5–10 in terms of variance.

Frankcombe et al. (2015) [see their figure 5(c)] and Trenary and DelSole (2016) previously noted a striking lack of internal AMO variance in CMIP5 model simulations. Mann et al. (2014) concluded that this property must imply that the observed multidecadal excursions of AMO, unmatched by most of the CMIP5 models, are largely externally forced (see also Mann and Emanuel, 2006). In this interpretation, more pronounced multidecadal undulations of the observed surface temperatures would be due to models’ underestimating the multidecadal component of the true forced climate response, while the true internal variability in observations would be consistent with the simulated internal variability. Indeed, Booth et al. (2012) proof-of-concept results demonstrated quantitative feasibility of this scenario; in their model (HadGEM2-ES), indirect aerosol effects drove multidecadal AMO variability in SST that closely matched observations after approximately 1950. Similarly, Golaz et al. (2013) documented large sensitivity of GFDL CM3 model to small changes in the cloud formulation parameters (via ensuing large changes in the magnitude of aerosol indirect effects on climate), implying that the magnitude of forced variations in simulated climates may be tunable. On the other hand, Zhang et al. (2013) identified a number of major discrepancies between the observed climate changes and those simulated in the HadGEM2-ES model, and attributed them to excessively strong aerosol effects. More generally, Stevens (2015) argued that the majority of CMIP5 models already have excessive aerosol forcing, which thus cannot fully explain the multidecadal downturns of the observed climate warming.

An alternative (or complementary) explanation of excessively damped multidecadal climate variability in CMIP5 models involves the possibility that these models misrepresent some of the dynamical feedbacks at work in the real climate system, in which case the model – data differences would reflect a lack or distortion of multidecadal internal dynamics in climate models. Multidecadal climate variations in the North Atlantic region have been associated with Atlantic Meridional Overturning Circulation (AMOC: Delworth et al., 1993; Timmermann et al., 1998; Delworth and Mann, 2000; Latif et al., 2004; Knight et al., 2005, 2006). The dynamics of the AMOC and North Atlantic SST signals are still not fully understood, and a large number of different theories are available (Delworth et al., 1993; Timmermann et al., 1998; Frankcombe et al., 2009, 2010; Frankcombe and Dijkstra, 2011; Clement et al., 2015; Trenary and DelSole, 2016). Meanwhile, the simulated signals exhibit a wide spread in their characteristic time scales and amplitudes across CMIP5 models (Zhang and Wang, 2013; Ba et al., 2014), and the climate response mostly confined to the North Atlantic region and its immediate surroundings (Enfield et al., 2001; Sutton and Hodson, 2005; Knight et al., 2006), perhaps with an in-phase (simultaneous) teleconnection to North Pacific (Kravtsov and Sannagle, 2008; DelSole et al., 2011; Wyatt and Peters, 2012; Kravtsov et al., 2014) or tropical Pacific (McGregor et al., 2014 and references therein).

Wyatt et al. (2012) and Kravtsov et al. (2014) argued that the spatiotemporal structure of observed multidecadal climate variability in the 20th century may be even more complex and involves hemispheric propagation of the AMO multidecadal signal, which they termed the ‘stadium wave’. They further noted the absence of the stadium wave in CMIP3 (Wyatt and Peters, 2012) and some of the CMIP5 models (Kravtsov et al., 2014) and argued it to be due to the models’ failing to transmit the simulated
AMO signal to the overlying atmosphere, which may in turn result from the insufficient atmospheric response to SST or sea-ice anomalies in model simulations; see Kushnir et al. (2002) and Wyatt and Curry (2014). The stadium wave — absent from CMIP models — may be one possible explanation of the discrepancies between the observed (semi-empirical) and CMIP5 simulated internal climate variability documented in the present study. We plan to address this hypothesis in a future work.

Finally, we venture to speculate about a hypothetical scenario in which climate models have a more pronounced multidecadal internal variability consistent with our semi-empirical estimates. In this case, a combination of a weaker forced warming trend and large multidecadal climate variability could easily produce synthetic realizations of climate variability as close to the observed variations as the current CMIP5 simulations dominated by the strong non-uniform forced trend. Were such a configuration possible, it could introduce significant upward revisions in the Mann et al.’s (2016) estimates of the likelihood of the recent warming to occur in the absence of anthropogenic effects.

To conclude, state-of-the-art climate models are characterized by a substantial model uncertainty, large sensitivity to aerosol and cloud parameterizations and a possible lack of feedbacks that could amplify multidecadal internal variability, which impedes clear attribution of the observed 20th century climate change.

Acknowledgements

We acknowledge the World Climate Research Programme’s Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modelling groups for producing and making available their model output. ERSST data were provided by NOAA: www.ncdc.noaa.gov/data-access/marineocean-data/extended-reconstructed-sea-surface-temperature-ersst-v3b, https://climatedataguide.ucar.edu/climate-data/naoa-20th-century-reanalysis-version-2-and-2c. Kaplan SST V2 data were provided by the NOAA/OFFice of Oceanic and Atmospheric Research/Earth System Research Laboratory Physical Sciences Division, Boulder, Colorado, USA: www.esrl.noaa.gov/psd. HadISST data were provided by the Met Office Hadley Centre: www.metoffice.gov.uk/hadobs. We also thank the authors of Steinman et al. (2015a) for making their data publicly available. We are grateful to three anonymous reviewers for their comments and suggestions, which helped us to improve the presentation. This research was supported by NSF grants OCE-1243158 and AGS-1408897. All data and MATLAB® scripts for this paper are available for downloading from http://atmo.math.uwm.edu: 8181/ → S. Kravtsov Data → KC2016.

Supporting information

The following supporting information is available as part of the online article:

Appendix S1. Supporting information.

References

Allen MR, Tett SFJ. 1999. Checking for model consistency in optimal fingerprinting. Clim Dyn 15: 419–434.
Ba J, Keeleyhides NS, Latif M, Park W, Ding H, Lohmann K, Mignot J, Menary M, Ottera OH, Wouters B, Melia DS, Oka A, Bellucci A, Volodin E. 2014. A multi-model comparison of Atlantic decadal variability. Clim Dyn 43: 2333–2348. https://doi.org/10.1007/s00382-014-2056-1.
Booth BBB, Dunstone NJ, Halloran PR, Andrews R, Bellouin N. 2012. Aerosols implicated as a prime driver of twentieth-century North Atlantic climate variability. Nature 484: 228–232.
Clement A, Bellomo K, Murphy L, Cane M, Maufrtsen T, Rådel G, Stevens B. 2015. The Atlantic Multidecadal Oscillation without a role for ocean circulation. Science 340: 320–324. https://doi.org/10.1126/science.aab3980.
Cook J, Nuccitelli D, Green SA, Richardson M, Winkler B, Painting R, Way R, Jacobs P, Skuce A. 2013. Quantifying the consensus on anthropogenic global warming in the scientific literature. Environ. Res. Lett. 8: 024024. https://doi.org/10.1088/1748-9326/8/2/024024.
Davy R, Esau I. 2016. Difference in the efficacy of climate forcings explained by variations in atmospheric boundary layer depth. Nat. Commun. 7: 11690. https://doi.org/10.1038/ncomms11690.
DelSole T. 2001. Optimally persistent patterns in time-varying fields. J. Atmos. Sci. 58: 1341–1356.
DelSole T. 2006. Low-frequency variations of surface temperature in observations and simulations. J. Clim. 19: 4487–4507.
DelSole T, Shukla J. 2010. Model fidelity versus skill in seasonal forecasting. J. Clim. 18: 4794–4806.
DelSole T, Tippett MK. 2007. Predictability: recent insights from information theory. Rev. Geophys. 45: RG4002. https://doi.org/10.1029/2006RG000202.
DelSole T, Tippett MK. 2009. Average predictability time: part II: seamless diagnosis of predictability on multiple time scales. J. Atmos. Sci. 66: 1188–1204.
DelSole T, Tippett MK, Shukla J. 2011. A significant component of unfurced multidecadal variability in the recent acceleration of global warming, J. Clim. 24: 909–926.
DelSole T, Jia L, Tippett MK. 2013. Decadal prediction of observed and simulated sea surface temperatures. Geophys. Res. Lett. 40: 2773–2778.
Deworh TL, Mann ME. 2000. Observed and simulated multidecadal variability in the Northern Hemisphere. Clim Dyn 16: 661–676.
Deworh TL, Manabe S, Stouffer R. 1993. Interdecadal variability of the thermohaline circulation in a coupled ocean–atmosphere model. J. Clim. 6: 1993–2001.
Elsner JB, Tsonis AA. 2006. Singular Spectrum Analysis: A New Tool in Time Series Analysis. Springer: New York. https://doi.org/10.1007/978-1-4757-2514-8, 164 pp.
Enfield DB, Mestas-Nuñez AM, Trimble PJ. 2001. The Atlantic multidecadal oscillation and its relation to rainfall and river flows in the continental US. Geophys. Res. Lett. 28: 277–280.
Evant AT, Bennetts KJ, Birmont DJ. 2013. The modification of sea surface temperature anomaly linear damping time scales by stratocumulus clouds. J. Clim. 26: 3619–3630.
Frankcombe LM, Dijkstra HA. 2011. The role of Atlantic–Arctic exchange in North Atlantic multidecadal climate variability. Geophys. Res. Lett. 38: L16603. https://doi.org/10.1029/2011GL048158.
Frankcombe LM, Dijkstra HA, Von der Heydt A. 2009. Noise-induced multidecadal variability in the North Atlantic: excitation of normal modes. J. Phys. Oceanogr. 39: 220–233.
Frankcombe LM, Von der Heydt A, Dijkstra HA. 2010. North Atlantic multidecadal climate variability: an investigation of dominant time scales and processes. J. Clim. 23: 3626–3638.
Frankcombe LM, England MH, Mann ME, Steinman BA. 2015. Separating internal variability from the externally forced climate response. J. Clim. 28: 8184–8202. https://doi.org/10.1175/JCLI-D-15-0069.1.
Ghil M, Vautard R. 1991. Interdecadal oscillations and the warming trend in global temperature time series. Nature 350: 324–327. https://doi.org/10.1038/350324a0.
Ghil M, Allen RM, Dettinger MD, Ide K, Kondrashov D, Mann ME, Robertson AW, Saunders A, Tian Y, Varadi F, Yiou P. 2002.
Wu Z, Huang NE, Wallace JM, Smoliak B, Chen X. 2011. On the time-varying trend in global-mean surface temperature. *Clim. Dyn.* **37**: 759–773. https://doi.org/10.1007/s00382-011-1128-8.

Wyatt MG, Curry JA. 2014. Role for Eurasian Arctic shelf sea ice in a secularly varying hemispheric climate signal during the 20th century. *Clim. Dyn.* **42**: 2763–2782. https://doi.org/10.1007/s00382-013-1950-2.

Wyatt MG, Peters JM. 2012. A secularly varying hemispheric climate-signal propagation previously detected in instrumental and proxy data not detected in CMIP3 database. *SpringerPlus* **1**: 68. https://doi.org/10.1186/2193-1801-1-68.

Wyatt M, Kravtsov S, Tsonis AA. 2012. Atlantic Multidecadal Oscillation and Northern Hemisphere’s climate variability. *Clim. Dyn.* **38**: 929–949. https://doi.org/10.1007/s00382-011-1071-8.

Xue Y, Smith TM, Reynolds RW. 2003. Interdecadal changes of 30-yr SST normals during 1871–2000. *J. Clim.* **16**: 1601–1612.

Zhang L, Wang C. 2013. Multidecadal North Atlantic sea surface temperature and Atlantic Meridional Overturning Circulation variability in CMIP5 historical simulations. *J. Geophys. Res. Oceans* **118**(10): 5772–5791. https://doi.org/10.1002/jgrc.20390.

Zhang R, Delworth TL, Held IM. 2007. Can the Atlantic Ocean drive the observed multidecadal variability in Northern Hemisphere mean temperature? *Geophys. Res. Lett.* **34**: L02709. https://doi.org/10.1029/2006GL028683.

Zhang R, Delworth TL, Sutton R, Hodson DLR, Dixon KW, Held IM, Krishnir Y, Marshall J, Ming Y, Msadek R, Robson J, Rosati AJ, Ting MF, Vecchi GA. 2013. Have aerosols caused the observed Atlantic Multidecadal Variability? *J. Atmos. Sci.* **70**: 1135–1144. https://doi.org/10.1175/JAS-D-12-0331.1.