On the Choice of Ensemble Mean for Estimating the Forced Signal in the Presence of Internal Variability

LEELA M. FRANKCOMBE AND MATTHEW H. ENGLAND

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

JULES B. KAJTAR

College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, United Kingdom

MICHAEL E. MANN

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

BYRON A. STEINMAN

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

(Manuscript received 3 October 2017, in final form 17 April 2018)

ABSTRACT

In this paper we examine various options for the calculation of the forced signal in climate model simulations, and the impact these choices have on the estimates of internal variability. We find that an ensemble mean of runs from a single climate model [a single model ensemble mean (SMEM)] provides a good estimate of the true forced signal even for models with very few ensemble members. In cases where only a single member is available for a given model, however, the SMEM from other models is in general out-performed by the scaled ensemble mean from all available climate model simulations [the multimodel ensemble mean (MMEM)]. The scaled MMEM may therefore be used as an estimate of the forced signal for observations. The MMEM method, however, leads to increasing errors further into the future, as the different rates of warming in the models causes their trajectories to diverge. We therefore apply the SMEM method to those models with a sufficient number of ensemble members to estimate the change in the amplitude of internal variability under a future forcing scenario. In line with previous results, we find that on average the surface air temperature variability decreases at higher latitudes, particularly over the ocean along the sea ice margins, while variability in precipitation increases on average, particularly at high latitudes. Variability in sea level pressure decreases on average in the Southern Hemisphere, while in the Northern Hemisphere there are regional differences.

1. Introduction

The climate we observe is made up of an externally forced component (dominated by the anthropogenic warming trend, interspersed with the volcanic signal) and a component due to the internal variability of the climate system. Despite the fact that internal variability and the forced signal are not necessarily separable, especially on regional scales (see, e.g., Otterå et al. 2010; Maher et al. 2015; Swingedouw et al. 2017), there are many analyses for which it is useful to study the internal variability and/or the forced signal in isolation, in so far as it is possible. Accordingly, there has been some discussion on the best way to achieve the separation of the two components. Previous methods include removing a linear trend (e.g., Wyatt et al. 2012; Chylek et al. 2014, among many others), removing the regression of the global mean from regional sea surface temperatures (SSTs; e.g., Trenberth and Shea 2006), removing the

Corresponding author: Leela M. Frankcombe, l.frankcombe@unsw.edu.au

DOI: 10.1175/JCLI-D-17-0662.1

regression of the global mean and an estimate of aerosol forcing (Mann and Emanuel 2006), and removing the simple mean of an ensemble of climate simulations (e.g., Knight 2009). However some of these methods, particularly linear detrending, have been shown to be inaccurate and may give rise to misleading results (Mann et al. 2014; Steinman et al. 2015a; Frankcombe et al. 2015).

One method of isolating the internal variability from the response to external forcing is to estimate the forced response using the average of an ensemble of simulations from a climate model. Following from our assumption of the separability of the forced and internal components, the phase, amplitude, and periodicity of internal variability are functions of the initial conditions only. Thus, given an increasingly large ensemble of simulations from the same model driven with identical external forcing, the ensemble average will converge to the true forced response. Once this forced response has been estimated, it can then be removed from individual simulations, and what remains is the internal variability. However, when making use of an ensemble such as phase 5 of the Coupled Model Intercomparison Project (CMIP5), which contains members from different climate models that will have slightly different responses to the same forcing, additional errors are introduced.

In applying the forced signal from the models to observations, even more potential errors are introduced, since the external forcings applied to the models are not necessarily correct and complete, in that the model responses to those external forcings will also contain errors. To partly ameliorate this error, the model-forced signal is scaled to match observations (to take into account potentially different rates of warming in the models and the real world). The remainder, after this scaled forced signal is subtracted from the observations, then provides an estimate of the observed internal variability (Steinman et al. 2015a; Frankcombe et al. 2015). There is, however, debate about the best way of constructing the ensemble mean of the climate models so as to minimize errors. Steinman et al. (2015a) used the multimodel ensemble mean (MMEM), constructed as the average of all the available CMIP5 models, as well as an MMEM from a subset of the CMIP5 models (those containing aerosol indirect effects). They also tested the effect of using a single-model ensemble mean (SMEM) from models with 10 or more members in their ensembles. In each case the estimate of the forced signal is scaled to match the observations or model results (the so-called scaled MMEM or scaled SMEM methods). The scaled MMEM method has been shown, in models, to be significantly better than linearly detrending or using an unscaled MMEM estimate for the forced signal (Frankcombe et al. 2015). Kravtsov et al. (2015), Kravtsov (2017), and Kravtsov and Callicutt (2017) claimed that the SMEM method, since it is constructed using only ensemble members from individual climate models, is a more accurate approach because it accounts for the differences in sensitivities of different climate models to the various types of external forcings. Steinman et al. (2015b) and Cheung et al. (2017a,b) maintained that the scaled MMEM is a more useful method in practice because of the limited number of ensemble members available to construct the SMEMs. Here we return to this question using synthetic data where the ‘‘forced’’ and ‘‘internal’’ components are known by construction, and take a more detailed look at the errors arising from each method, as well as the range of estimates of internal variability obtained using the different estimates of the forced signal. We then apply the method to a future

TABLE 1. Table of CMIP5 models used. The number of ensemble members available for the historical + RCP8.5 scenario are listed in the second column, and the length of the control run in years is listed in the third column. ACCESS1.0, ACCESS1.3, BCC_CSM1.1, BCC-CSM1.1(m), CESM1(BGC), CMCC-CM, CMCC-CMS, GFDL CM3, GFDL-ESM2G, GFDL-ESM2M, GISS-E2-H-CC, GISS-E2-R-CC, HadGEM2-AO, HadGEM2-CC, INM-CM4.0, IPSL-CM5A-MR, IPSL-CM5B-LR, MIROC-ESM, MRI-CGCM3, MRI-ESM1, NorESM1-M, and NorESM1-ME each had only one ensemble member available. EC-EARTH had six ensemble members available and MIROC5 had five, but the time series did not extend to the end of the RCP scenario. These models were therefore used in the calculation of the MMEM but not for estimates of future variability and are not listed in the table. (Expansions of acronyms can be found online at http://www.ametsoc.org/PubsAcronymList.)

Model name	Historical + RCP8.5	Control run length
CanESM2	5	996
CCSM4	6	1051[a]
CESM1(CAM5)	3	319
CNRM-CM5	5[b]	850[c]
CSIRO Mk3.6.0	10	500
FGOALS-s2	3	501
FIO-ESM	3[d]	800
GISS-E2-H (p1)	2[e]	540
GISS-E2-H (p3)	2	531
GISS-E2-R (p1)	2	550
GISS-E2-R (p3)	2[e]	531
HadGEM2-ES	4	575
IPSL-CM5A-LR	4	1000
MPI-ESM-LR	3	1000

[a] 501 years of control run data were available for precipitation.
[b] No data were available for precipitation.
[c] 800 years of control run data were available for SLP.
[d] Only two ensemble members were available for SLP and none for precipitation.
[e] No ensemble members were available for SAT.

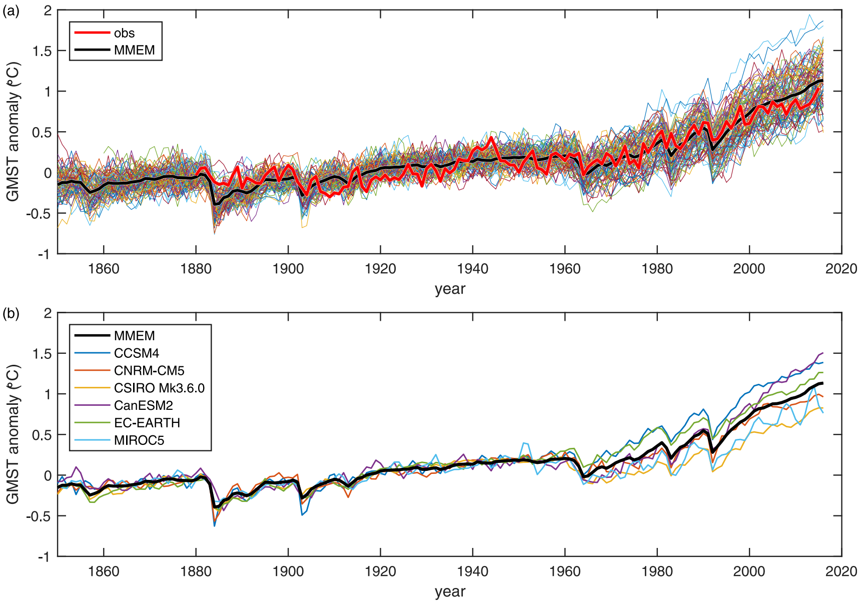

FIG. 1. (a) GMST anomalies from CMIP5 models (thin colored lines), along with the MMEM (black) and the observations (red). (b) MMEM (black) and SMEMs (colored lines) for GMST anomalies from CMIP5 models. Anomalies are calculated relative to the mean over the period 1880–1960.

scenario from the CMIP5 archive to obtain estimates of internal variability under increased anthropogenic forcing.

2. Method

We use SSTs, surface air temperatures (SATs), sea level pressure (SLP), and precipitation data from the preindustrial control runs, the historical runs, and future scenario (RCP4.5 and RCP8.5) runs from CMIP5 (Taylor et al. 2012). The CMIP5 models used are listed in Table 1. For observations we use monthly SST from the HadISST1 dataset (Rayner et al. 2003) between 1870 and 2015, and global surface temperatures from GISTEMP (Hansen et al. 2010) between 1880 and 2015. For comparison with observations, the CMIP5 historical runs were extended from 2005 to 2015 using RCP8.5. Note that we use model SATs rather than blended SSTs and SATs (Cowtan et al. 2015); however, testing showed that use of a blended product does not alter the conclusions. Likewise, model drift in the control run data was not corrected for, since detrending the control runs changed the variance by less than 0.01% per 100 yr on average. Smoothed time series are calculated using an adaptive low-pass filter (Mann 2008).

There are various methods of constructing the ensemble mean from the CMIP5 ensemble to take into account model independence and/or performance (see, e.g., Haughton et al. 2015); however, in this idealized study we consider a simple mean of all available ensemble members (here called the MMEM) as our first estimate of the forced signal. Averaging simulations from each model and then averaging over all the models does not alter the conclusions. It is important to note that there is a distinction between the confidence interval of the MMEM and the potential difference between the MMEM and the true forced signal. The MMEM is calculated from a large number of ensemble members and thus has a narrow confidence interval, as shown by bootstrap resampling in Steinman et al. (2015a) and by the range of individual estimates of the internal variability in Steinman et al. (2015b). On the other hand, with no further information we cannot tell whether the MMEM thus calculated is an accurate representation of the observed forced signal.

To estimate potential errors in our methods for assessing internal variability, we therefore use a large ensemble of synthetic time series of global-mean surface air temperature (GMST), where the forced and internal variability components are known. These synthetic time series are designed to approximately resemble the CMIP5 ensemble. First the internal variability of the CMIP5 ensemble is characterized by removing the scaled MMEM from each member of the historical ensemble, then calculating the autocorrelation and amplitude of variability

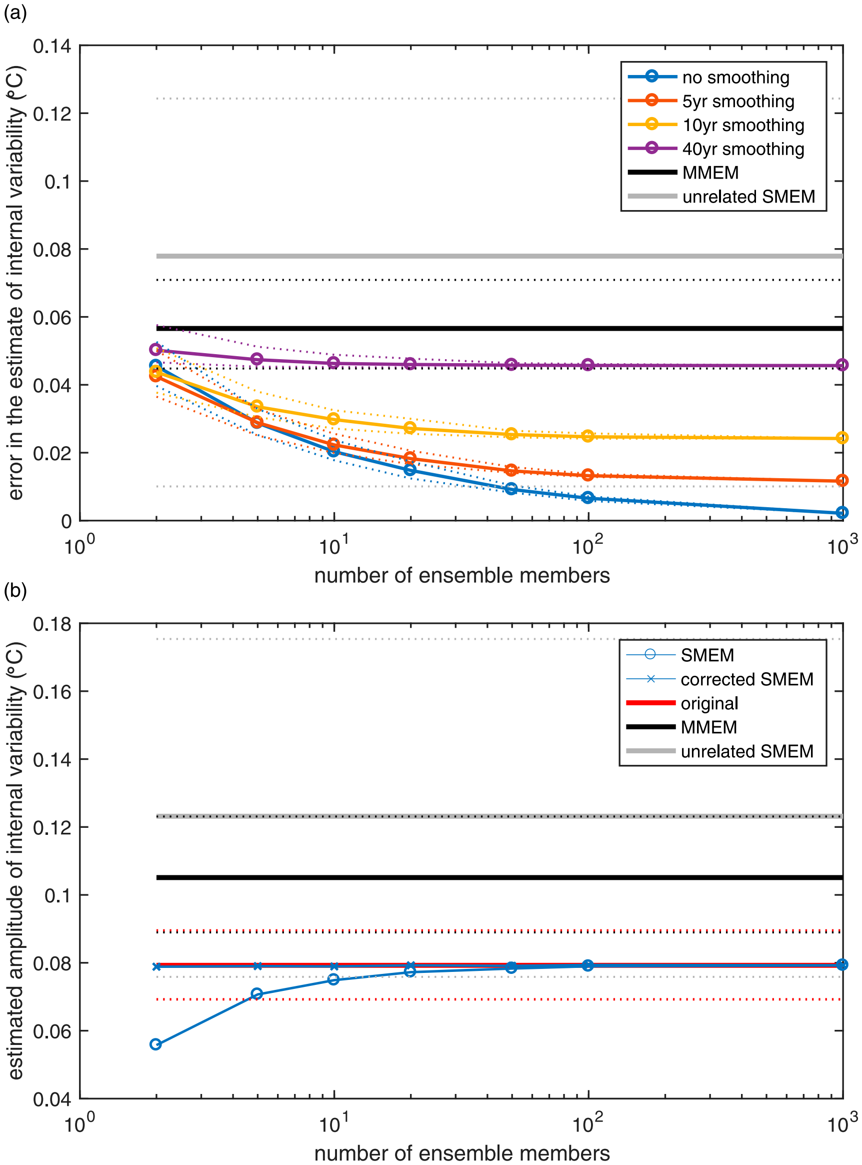

FIG. 2. Error in SMEM and MMEM methods for a synthetic ensemble representing GMST. Colored curves show (a) the error of the estimated time series of internal variability and (b) the standard deviation of the estimated time series of internal variability, where the variability is estimated using the SMEM method. The dependence on the number of ensemble members is shown on the *x* axis. For comparison, the black line shows the median error using the MMEM and the gray line shows the median error when applying the SMEM method to the whole ensemble (including time series generated using unrelated SMEMs). Solid lines show the median value; dotted lines show the 5th and 95th percentiles. In (a), the colors represent different smoothing time scales applied to the SMEM. The red line in (b) shows the amplitude of the original time series. In (b), only the results for the case where no smoothing is applied to the SMEM are shown, since the amplitude correction is not valid when smoothing is applied.

(as the standard deviation) of the resulting time series. The synthetic time series of GMST are then generated as 165-yr-long (the same length as the CMIP5 historical runs) time series of red noise, scaled by the average autocorrelation and amplitude of the internal variability estimated from the CMIP5 models. These synthetic time series of internal variability are then converted into time series of historical variability by adding either the MMEM or one of the SMEMs. Thus, the ''forced'' and ''internal'' components are known by construction. The construction and use of this type of synthetic time series is further described by Frankcombe et al. (2015).

When examining model variability under future forcing scenarios, we use several different indices. The Atlantic multidecadal oscillation (AMO) index is defined as the average SST in the region 0°–60°N, 5°–75°W, and the Pacific multidecadal oscillation (PMO) is defined as the average SST in the region 0°–60°N, 120°E–100°W. For the interdecadal Pacific oscillation (IPO), we use the tripole index of Henley et al. (2015). For ENSO, we use two indices in order to capture potential shifts in the location of ENSO variability in the future—the Niño-1.2 (SST in the region 0°–10°S, 90°–80°W) and Niño-3.4 (SST in the region 5°S–5°N, 170°–120°W) indices. For future variability of SLP, we look at changes in the southern annular mode (SAM) index, calculated as the difference between the normalized SLP at 40° and 65°S, and the Southern Oscillation index (SOI), which is defined as the pressure difference between normalized SLP at the model grid points closest to Tahiti and Darwin. These indices were chosen because of their use in past studies or because there is considerable interest in the future behavior of the modes of variability that they represent.

3. Results

Figure 1a shows the observed GMST index (red) along with the raw indices from the CMIP5 models (colors) and the MMEM (black) calculated as the average of all the ensemble members. Figure 1b shows the MMEM and six different SMEMs from CMIP5 models with five or more available ensemble members. We can see that the MMEM is smoother than the individual SMEMs, as the MMEM is constructed from a larger number of ensemble members and therefore the internal variability is more effectively averaged out. There is also considerable spread between the different SMEMs toward the end of the time series, as the different modeled rates of warming relative to the reference period become apparent.

a. *How many ensemble members are required to accurately estimate the forced signal?*

First we investigate the number of ensemble members from a single model that are required to accurately estimate the forced signal. This will allow us to judge whether it is viable to use the small single-model ensembles from CMIP5 to estimate internal variability or whether the residual forced signal that remains after removal of the ensemble mean is so large that any estimates are meaningless. Using our synthetic single-model ensemble, an estimate of the forced signal is calculated from a specified number of ensemble members. This estimated forced signal may also be smoothed to remove some of the unwanted residual variability. For example, Kravtsov et al. (2015) used a 5-yr smoothing window to calculate estimates of the forced signal for single-model ensembles. An estimate of the internal variability component is obtained by subtracting the estimated forced signal from the time series, and this estimate of the internal variability is then compared to the known internal variability of the original time series. The error is calculated as the square root of the sum of the squared differences at every time step between the original and estimated time series of internal variability. Figure 2a shows this error in the estimate of the internal variability for different numbers of ensemble members (on the *x* axis) and different values of the smoothing (different colors). Results using surrogates based on the GMST are shown in Fig. 2; results for the AMO and PMO are similar. We can see that the error decreases as the number of ensemble members increases, as expected. For very small ensemble sizes, less than about 10 members, some smoothing may slightly reduce the error. For larger ensembles, particularly for large smoothing windows, the smoothing becomes counterproductive. Out of 38 CMIP5 models included in this study, all but one have fewer than 10 ensemble

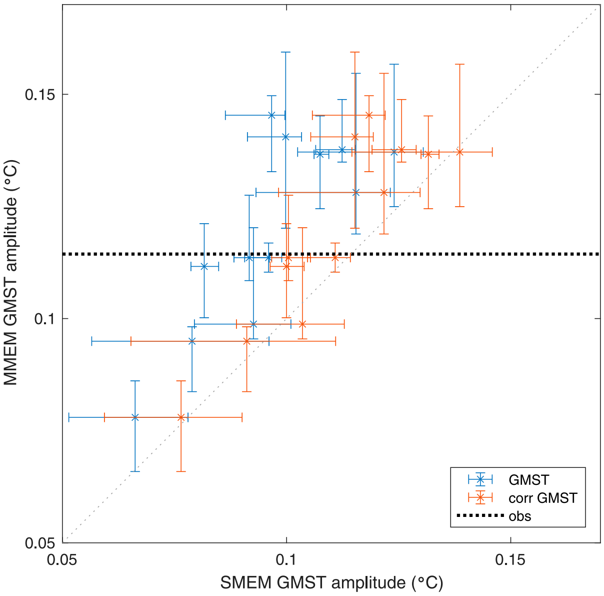

FIG. 3. Amplitudes for variability of GMST in the CMIP5 ensemble estimated using the MMEM method and the SMEM method with and without the correction factor applied. The error bars span the 5th–95th percentiles of the ensemble spread for each model. The observed value is shown as the horizontal dashed line.

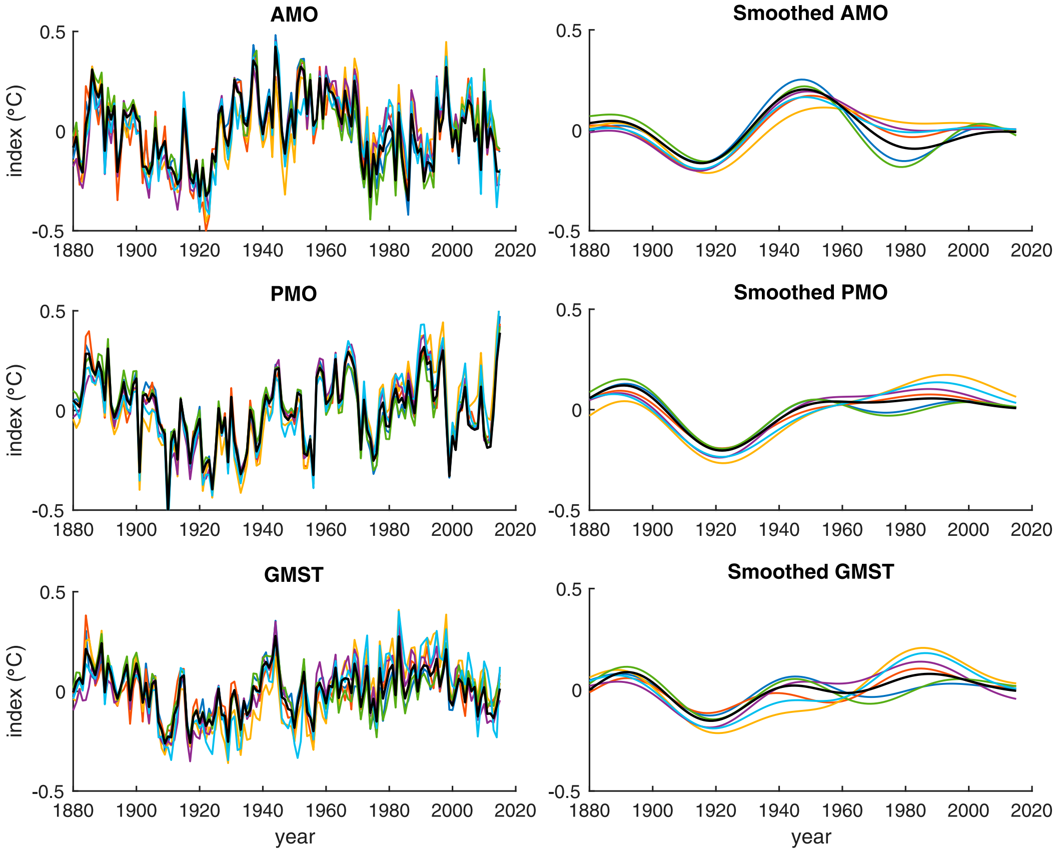

FIG. 4. The range of estimates for the observed (top) AMO, (middle) PMO, and (bottom) GMST indices obtained using the different SMEM estimates of the forced signal (using models with five or more ensemble members; colored lines). The MMEM estimate of each index is shown in black. Plots on the left show the annual data and plots on the right show the annual data smoothed with a 40-yr low-pass filter.

members, so in this case either no smoothing or smoothing with a small window gives the best result [e.g., the 5 years used by Kravtsov et al. (2015) is a sensible choice]. Note that there are slight differences between the effectiveness of smoothing for different indices. For example, for extremely small ensemble sizes, 40-yr smoothing gives slightly lower errors than no smoothing for the AMO, but not for the GMST. The effectiveness of smoothing is therefore dependent on both the ensemble size and the index being analyzed. In addition, the smoothing has an unphysical effect on the volcanic forced signal, and therefore we do not use smoothing of the forced signal any further in this analysis.

In cases of very small ensembles, such as those for many models in the CMIP5 archive, does it still make sense to use an SMEM, or would using the scaled MMEM instead be more accurate? Out of our ensemble of 38 CMIP5 models, 32 have fewer than five members in their ensemble (24 have just one member, and thus no SMEM can be calculated for these). To test the impact of using a small-ensemble SMEM versus the MMEM, we construct another ensemble of synthetic GMST data, this time using six different SMEMs (from each of the six CMIP5 models with five or more ensemble members). For each of these six sets of synthetic data we use the scaled MMEM as well as the six SMEMs (one related and five unrelated) to make seven sets of estimates of the internal variability. The errors in the time series of the internal variability thus obtained are plotted in Fig. 2a with the MMEM estimate in black and the SMEM estimate in gray. The median error for the SMEM-based estimates of the variability is higher than the MMEM-based estimate. These results show that while using an SMEM is more internally consistent for an individual model, the mean bias is potentially much larger than the MMEM method when applied to an unrelated model. This does not rule out a particular SMEM representing the forced signal better than the MMEM (e.g., in closely related models). However, we do not necessarily know which SMEM to use, particularly for observations, which may be considered as a model with one ensemble member. The MMEM

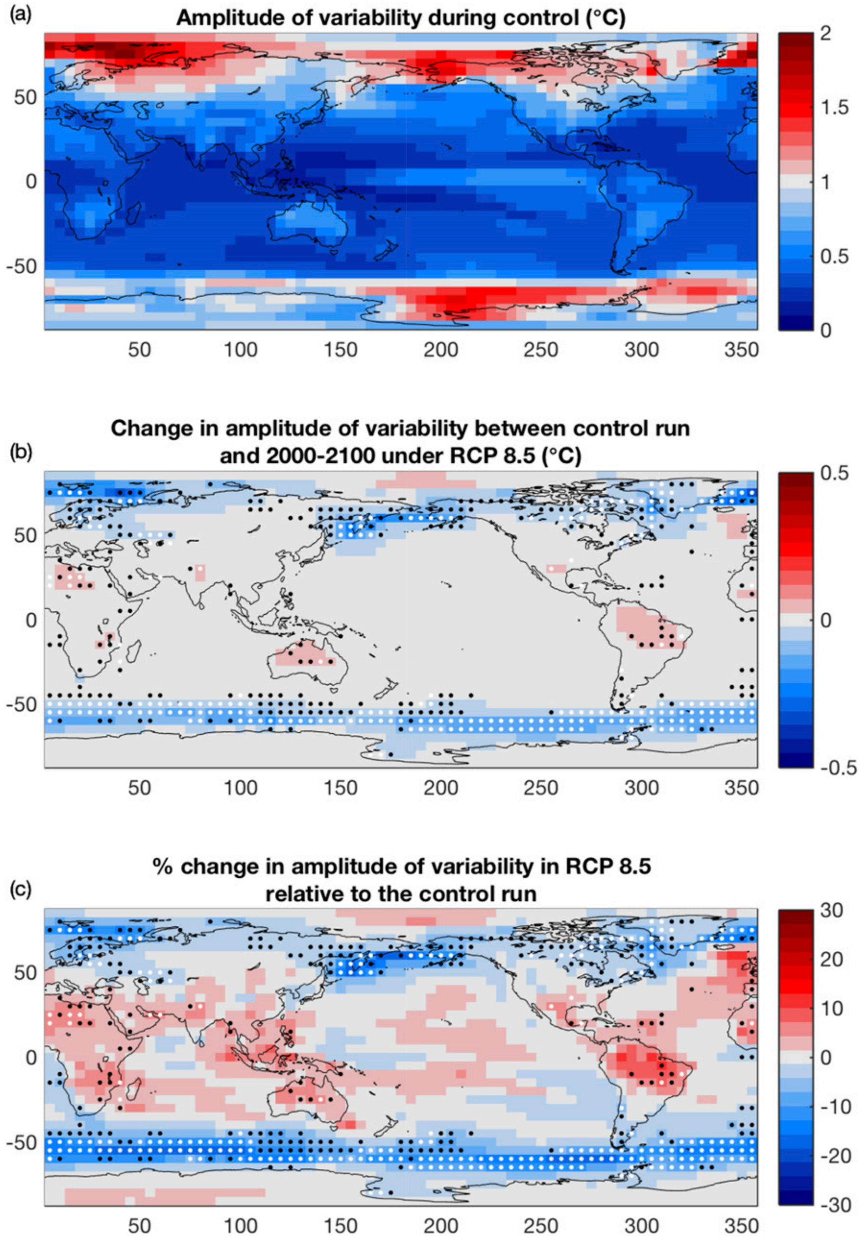
(a)
Amplitude of variability during control (°C)
50
0
-50
50
100
150
200
250
300
350
2
1.5
1
0.5
0
(b)
Change in amplitude of variability between control run and 2000-2100 under RCP 8.5 (°C)
50
0
-50
50
100
150
200
250
300
350
0.5
0
-0.5
(c)
% change in amplitude of variability in RCP 8.5 relative to the control run
50
0
-50
50
100
150
200
250
300
350
30
20
10
0
-10
-20
-30

FIG. 5. (a) Ensemble average spatial pattern of unsmoothed annual-mean SAT variability during the CMIP5 control runs, (b) change in amplitude between the control run and RCP8.5 over the period 2000–2100, and (c) percentage change in amplitude between the control run and RCP8.5 over the period 2000–2100. Black stippling shows where two-thirds of the models agree on the sign of the change, and white stippling shows where 90% of the models agree.

method is therefore a viable option for estimating the forced signal in cases where no SMEM is available.

One drawback of the SMEM method is that most of the models in the CMIP5 archive have very few ensemble members, which limits our ability to accurately estimate the forced response for those models. As we have seen in Fig. 2a, this can lead to significant errors in the estimation of the time series of internal variability for small ensembles. If the amplitude of the variability is calculated directly as the variance (or standard deviation) of each time series, this will result in large errors in the estimated amplitude of internal variability. However, following Olonscheck and Notz (2017), if the variance of the ensemble is calculated at every time step and then averaged over the length of the time series, rather than being calculated for each individual time

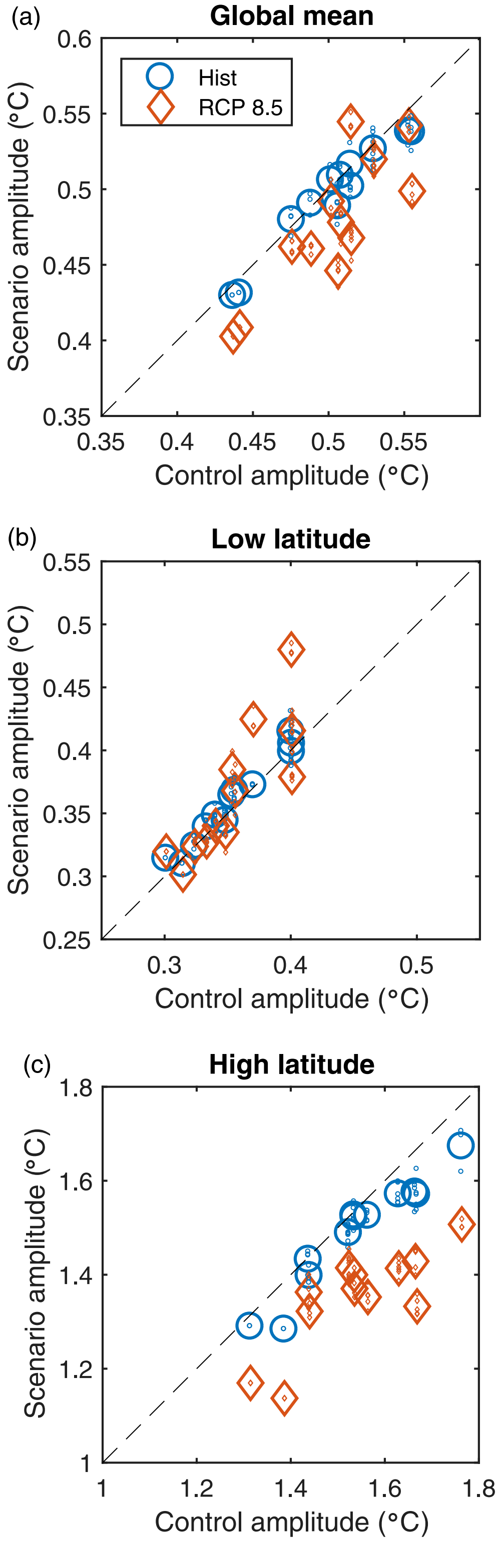

FIG. 6. Scatterplot of mean amplitude of unsmoothed annual-mean SAT variability in the control run compared to a future scenario for (a) the global mean, (b) low latitudes (40°S–40°N), and (c) high latitudes (poleward of 40°N and 40°S). Small symbols show the individual ensemble members while large symbols are the ensemble mean values. "Hist" covers the period 1900–2000 while "RCP 8.5" covers the period 2000–2100. Points on the dashed 1:1 line show that no change in the amplitude of the variability occurred between the control run and the historical or RCP scenario. Points below (above) the 1:1 line show that variability has decreased (increased) in the historical or RCP scenario compared to the control run. When only two ensemble members are available, both ensemble members will have the same amplitude of variability because of the method used to estimate the forced signal.

series and then averaged over the ensemble, we obtain an accurate estimate of the amplitude of the variability, as shown in Fig. 2b for the synthetic GMST data. This corresponds to a correction factor of $\sqrt{N/(N-1)}$, where N is the number of ensemble members. Thus an accurate estimate of the amplitude of the variability can be obtained with an ensemble of only two members. All amplitudes of variability calculated using the SMEM method from here onward include the small ensemble size correction from Olonscheck and Notz (2017). It must be noted, however, that this correction is applied only to the estimated amplitudes; it does not correct the estimated time series themselves. Accurate estimates of the time series of the variability still requires a larger ensemble, as shown in Fig. 2a.

In Fig. 3 we demonstrate the differences between the SMEM and MMEM estimates of the amplitude of the internal variability by applying the methods to the CMIP5 model ensemble. This figure compares the MMEM estimates of the variability in the historical simulations with the SMEM estimates of the variability (both uncorrected and corrected) for all models with two or more ensemble members. The MMEM method generally leads to higher estimates of the amplitude of the variability compared to the SMEM method, since the error in the estimate of the forced signal in the MMEM method will appear as additional variability, as discussed by Frankcombe et al. (2015) and Kravtsov and Callicutt (2017). The MMEM-based estimate of the amplitude of GMST variability in observations is shown as the dashed black line. Since observations may be considered as an ensemble with one member, it is not possible to calculate an SMEM-based estimate of the amplitude of the variability.

b. Application to observations

In application to the real world, the time series of observations may be treated as an ensemble with one member. There is no reason to assume that any one CMIP5 model more accurately estimates the real forced signal than the CMIP5 ensemble mean; therefore the

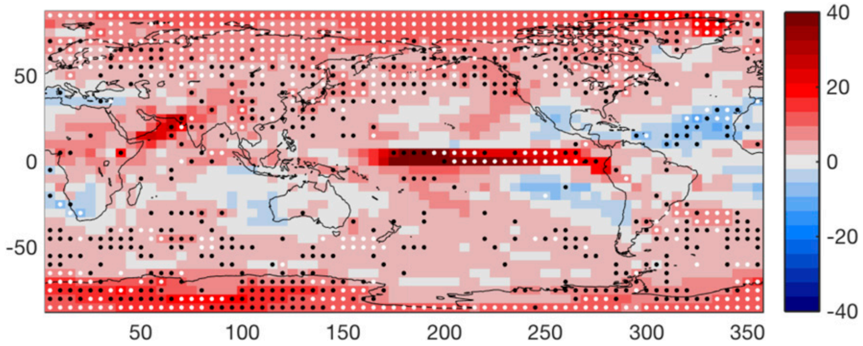

FIG. 7. Percentage change in amplitude of variability in unsmoothed annual-mean precipitation between the control run and RCP8.5 over the period 2000–2100. Black stippling shows where two-thirds of the models agree on the sign of the change, and white stippling shows where 90% of the models agree.

MMEM method remains a logical first choice for comparisons with observations. Figure 4 shows the estimates of the internal variability component of the observed AMO, PMO, and GMST indices calculated using the scaled MMEM (in black) and the six different scaled SMEMs. The choice of SMEM can make a considerable difference, particularly toward the end of the time series, because of the differing rates of warming of the individual models used to construct each SMEM. As discussed earlier, the use of an SMEM to estimate the forced signal in observations (which may be considered to be an unrelated model with one ensemble member) can potentially introduce larger biases in the estimates of internal variability than using the scaled MMEM.

The difference in these estimates of the internal variability highlights the importance of obtaining accurate estimates of the forced signal in order to correctly partition the observed signal into forced and internal components. For example, in Fig. 4, a majority of the estimates of the AMO index (including the MMEM estimate) show that the index was increasing and then levelled off toward the end of the time series; however, there is a large range of estimated amplitudes of the AMO index. All estimates of the PMO show decreases in the last one to two decades, as do most, but not all, of the estimates of the internal component of the GMST.

c. *Estimates of amplitudes of variability into the future*

One drawback of the MMEM method is that the errors in the estimate of the forced signal (and thus also in the estimate of the internal variability) increase markedly into the future, as the differences in the rates of warming between the models become increasingly important. The SMEM method, since it treats the different models separately, does not suffer from this problem, at least when calculating the amplitude of the variability. Thus we can use the SMEM method to obtain model estimates of the amplitude of internal variability further into the future for individual models with sufficiently large ensembles, and compare those future estimates to current or past variability.

Figure 5 shows the spatial pattern of the amplitude of variability of SAT during control runs as well as the change in the variability between the control run and the period 2000–2100 under RCP8.5, using the SMEM method. To obtain these patterns the standard deviation of SAT was calculated at each grid point over the specified period in each of the 13 models that have two or more ensemble members with the requisite data, and then the results of all the models were averaged. Figure 6 shows the amplitude of variability of annual SAT averaged over the globe as well as divided into low- and high-latitude bands for the control runs compared to the historical runs and the RCP8.5 scenario. In 11 of the 13 models, the globally averaged SAT variability decreases in the future. There is little agreement between the models on the spatial pattern of the change, especially at low latitudes, where most models show on average no change while a few models show a large increase in the amplitude of variability from the control run to the RCP8.5 scenario. At higher latitudes, however, the results are more consistent, with all the models showing a decrease in SAT variability, particularly over the ocean along the sea ice margins. This decrease in SAT variability over the ocean in a band at high latitudes exists at lower frequencies as well (using time series smoothed with 5- and 40-yr filters). Note that when only two ensemble members are available, both ensemble members will have, by definition, the same amplitude of variability as estimated by the SMEM method. This is because in these cases the SMEM is constructed using only two time series, and therefore the two time series of

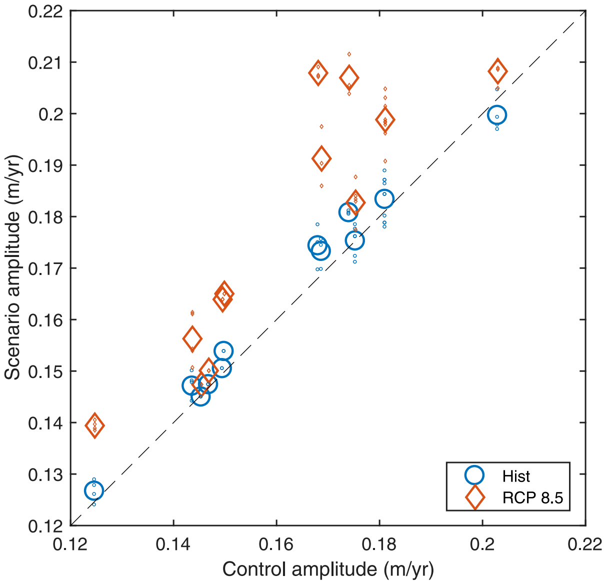

FIG. 8. Scatterplot of global-mean amplitude of variability in unsmoothed annual-mean precipitation in the control run compared to scenario runs. Symbols are as in Fig. 6.

variability that result when subtracting the SMEM are perfectly anticorrelated.

These results are largely similar to results from studies looking at the changes in higher-frequency variability (Huntingford et al. 2013; Screen 2014; Holmes et al. 2016; Olonscheck and Notz 2017, etc.). For example, Huntingford et al. (2013) found that overall variability will decrease under high greenhouse gas forcing scenarios, with the largest decreases in a band at around 50°–70° in each hemisphere. It also confirms the results of Olonscheck and Notz (2017) and Brown et al. (2017), who found similar patterns of changes in variability and associated the decrease at high latitudes with the loss of sea ice volume and the accompanying reduction in variability of albedo and increase in surface heat capacity of the open ocean compared to sea ice, while the increase in variability over land at low latitudes was linked to the decreasing availability of surface moisture as the mean temperature increases.

Using the same method we can also calculate the amplitude of simulated indices of variability such as the AMO, PMO, IPO, and ENSO in control, historical, and future scenarios. For annually averaged as well as 5- and 40-yr smoothed indices, most models do not show robust changes in the amplitude of the variability (i.e., a change that is larger than the spread between the different ensemble members of each model), and for models that do show robust changes, there is no agreement on the sign of the change in variability, as has been discussed in the literature for ENSO (e.g., Taschetto et al. 2014).

The model results also show that there is an overall increase in the amplitude of the variability of precipitation in RCP8.5 compared to the control runs, as shown in Fig. 7 for the spatial pattern of the percentage change in annual-mean precipitation, and Fig. 8 for the change in global mean. This is in broad agreement with past studies showing an increase in both wet and dry extremes in future scenarios (e.g., Sillmann et al. 2013; Alexander and Arblaster 2017). The largest regional change is in the equatorial Pacific; however, this may be an artifact of the shifting of the double ITCZ, which appears in many of the models. The most robust changes across the ensemble occur at high latitudes, particularly over the Northern Hemisphere where there are increases in variability of up to 20%. The large apparent magnitude of the change in the amplitude of the variability is due to the low mean precipitation in this region under preindustrial conditions and accompanies the well-known increase in mean precipitation over the Arctic in both observations and models under increasing greenhouse forcing (e.g., Kattsov and Walsh 2000;

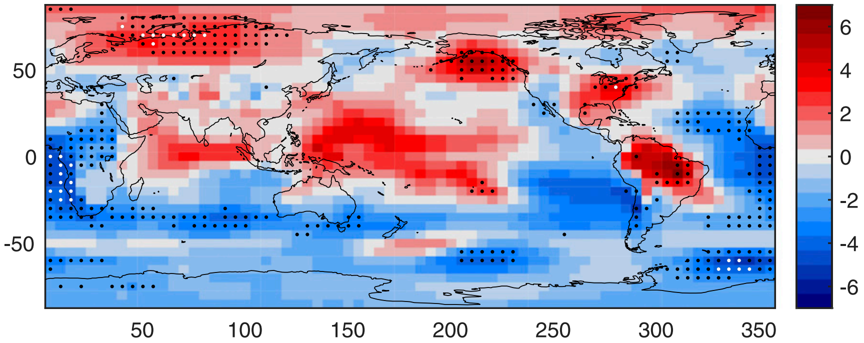

FIG. 9. Percentage change in amplitude of unsmoothed annual-mean SLP variability between the control run and RCP8.5 over the period 2000–2100. Black stippling shows where two-thirds of the models agree on the sign of the change, and white stippling shows where 90% of the models agree.

Kattsov et al. 2007). There is also an area of increased rainfall variability in the Arabian Sea, indicating possible changes in monsoonal rainfall in the region. There are regions of decreasing rainfall variability, mostly over the ocean basins at midlatitudes, although these may not be robust apart from the midlatitude North Atlantic where there is some agreement between the models. At lower frequencies the model average of the spatial patterns of the change in precipitation variability is similar to the annual-mean variability; however, there is less agreement between the models.

The spatial pattern of the change in the amplitude of variability in SLP shows an average increase in amplitude in the Northern Hemisphere and decrease in the Southern Hemisphere (Fig. 9). There is not a great deal of agreement between the models apart from in a few centers of action; however, hemispheric means (Fig. 10) show that most models predict a small decrease in the amplitude of variability in the Southern Hemisphere while results for the Northern Hemisphere are mixed. This may be related to the finding from Barnes and Polvani (2013) that under future climates in both the Southern Hemisphere and the North Atlantic the midlatitude jet moves poleward and exhibits less meridional shifting, while in the North Pacific the jet exhibits more meridional shifts.

In line with the prediction of decreased SLP variability in the Southern Hemisphere is the finding that there is a decrease in the amplitude of variability on annual and 5-yr time scales for the SAM (Fig. 11). No robust changes are seen for the SOI, which agrees with the lack of model agreement on the future behavior of the SAT-based ENSO index.

4. Conclusions

The issue of how best to separate internal variability from the forced signal is a nuanced one. It has previously been shown (Mann et al. 2014; Steinman et al. 2015a,b; Frankcombe et al. 2015) that the heretofore commonly used method of linear detrending introduces large errors and that the removal of a scaled ensemble mean is a more accurate method. The discussion has now moved to the choice of construction of that ensemble mean. In this paper we have shown that where multiple ensemble members from a model are available, a good estimate of the forced signal for that model can be calculated using the single-model ensemble mean (SMEM) method. The amplitude of internal variability thus calculated can be corrected to take into account the small ensemble size; however, the time series themselves will still contain some error. Where only a single time series is available (as is the case for observations, as well as a significant

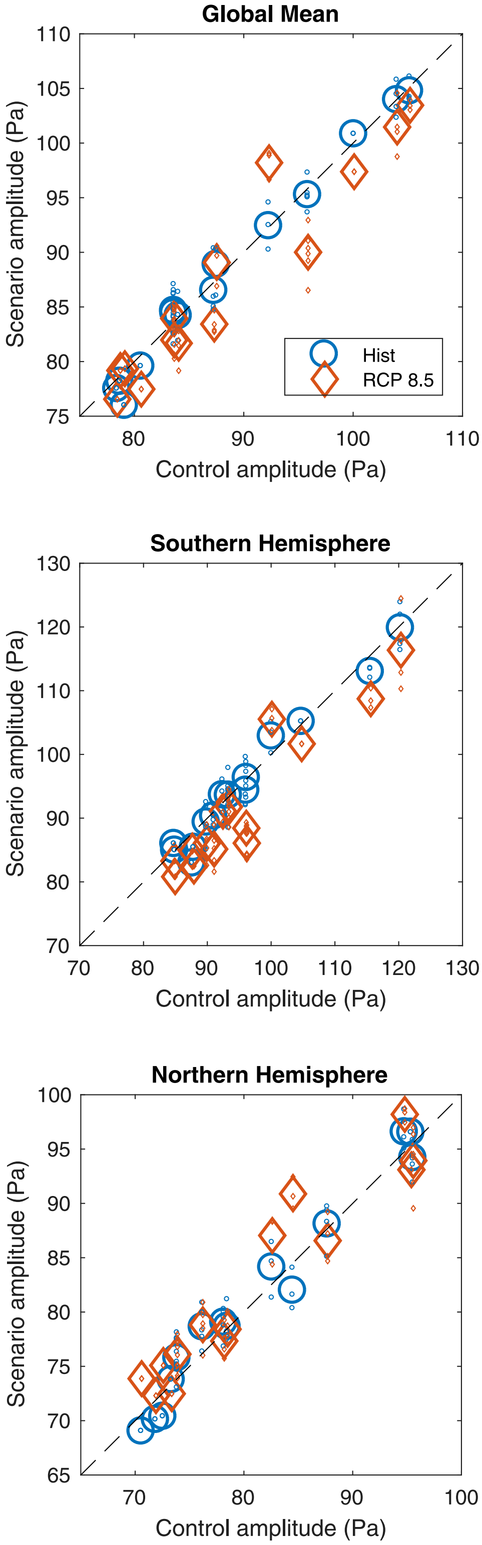

FIG. 10. Scatterplot of mean amplitude of unsmoothed annual-mean SLP variability in the control run compared to scenario runs for the (a) global mean, (b) Southern Hemisphere, and (c) Northern Hemisphere. Symbols are as in Fig. 6.

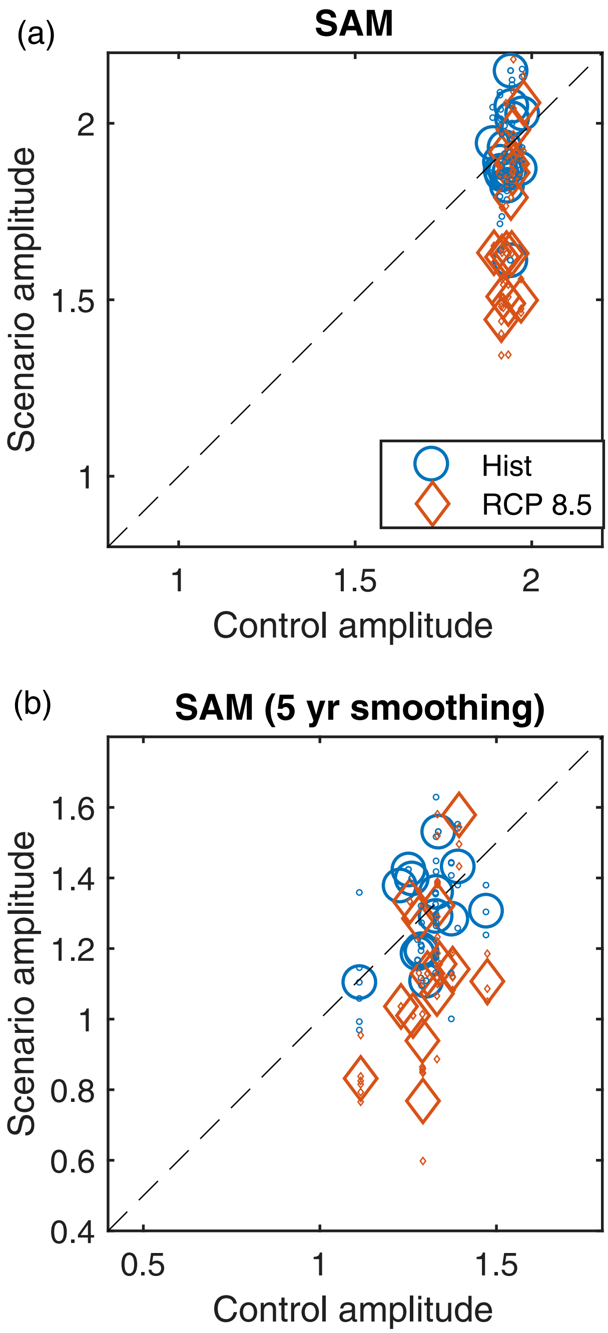

FIG. 11. Scatterplot of mean amplitude of variability of the SAM in the control run compared to scenario runs for (a) annual variability and (b) 5-yr variability. Symbols are as in Fig. 6.

proportion of the CMIP5 archive), the scaled multimodel ensemble mean (MMEM) estimate of the forced signal gives, on average, smaller errors than an estimate of the forced signal from an unrelated model's SMEM.

As an illustration of the use of the SMEM method we have calculated the change in the amplitude of annual-mean SAT, precipitation, and SLP variability in future climate under RCP8.5 for the models with multiple ensemble members for this scenario. We confirm the results of Huntingford et al. (2013), Olonscheck and Notz (2017), and Brown et al. (2017) (among others) that there are robust decreases in the variability of SAT along the sea ice margins in both hemispheres. We also see robust increases in the variability of precipitation, particularly at high latitudes [as follows from the results of Kattsov et al. (2007) that the mean precipitation at high latitudes increases under anthropogenic warming], and less robust but potentially interesting hemisphere-wide changes in SLP variability.

In summary, we find that both the MMEM and SMEM methods are useful, and to some extent, complementary. The SMEM method is the most accurate when applied to each model individually, especially for future scenarios. However, an SMEM cannot be calculated for observations, and while applying an SMEM from one of the models may result in a more accurate estimate of the observed forced signal than using the MMEM, it is impossible to know at this stage which model is the most correct one to use. Our results therefore indicate that the scaled MMEM method remains the most sensible choice for the estimation of the observed forced signal.

Acknowledgments. This work was supported by the Australian Research Council (ARC) through grants to L. M. F. (DE170100367) and to M. H. E. through the ARC Centre of Excellence in Climate System Science (CE110001028). J. B. K. is supported by the Natural Environment Research Council (Grant NE/N005783/1). B. A. S. was supported by the U.S. National Science Foundation (EAR-1447048). We 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) and GISTEMP by NASA Goddard Institute for Space Studies (data.giss.nasa.gov/gistemp).

REFERENCES

Alexander, L. V., and J. M. Arblaster, 2017: Historical and projected trends in temperature and precipitation extremes in Australia in observations and CMIP5. *Wea. Climate Extremes*, **15**, 34–56, https://doi.org/10.1016/j.wace.2017.02.001.

Barnes, E. A., and L. Polvani, 2013: Response of the midlatitude jets, and of their variability, to increased greenhouse gases in the CMIP5 models. *J. Climate*, **26**, 7117–7135, https://doi.org/10.1175/JCLI-D-12-00536.1.

Brown, P. T., Y. Ming, W. Li, and S. A. Hill, 2017: Change in the magnitude and mechanisms of global temperature variability with warming. *Nat. Climate Change*, **7**, 743–748, https://doi.org/10.1038/nclimate3381.

Cheung, A. H., M. E. Mann, B. A. Steinman, L. M. Frankcombe, M. H. England, and S. K. Miller, 2017a: Comparison of low-frequency internal climate variability in CMIP5 models and observations. *J. Climate*, **30**, 4763–4776, https://doi.org/10.1175/JCLI-D-16-0712.1.

——, ——, ——, ——, ——, and ——, 2017b: Reply to "Comment on 'Comparison of low-frequency internal climate variability in CMIP5 models and observations.'" *J. Climate*, **30**, 9773–9782, https://doi.org/10.1175/JCLI-D-17-0531.1.

Chylek, P., J. D. Klett, G. Lesins, M. K. Dubey, and N. Hengartner, 2014: The Atlantic multidecadal oscillation as a dominant factor of oceanic influence on climate. *Geophys. Res. Lett.*, **41**, 1689–1697, https://doi.org/10.1002/2014GL059274.

Cowtan, K., and Coauthors, 2015: Robust comparison of climate models with observations using blended land air and ocean sea surface temperatures. *Geophys. Res. Lett.*, **42**, 6526–6534, https://doi.org/10.1002/2015GL064888.

Frankcombe, L. M., M. H. England, M. E. Mann, and B. A. Steinman, 2015: Separating internal variability from the externally forced climate response. *J. Climate*, **28**, 8184–8202, https://doi.org/10.1175/JCLI-D-15-0069.1.

Hansen, J., R. Ruedy, M. Sato, and K. Lo, 2010: Global surface temperature change. *Rev. Geophys.*, **48**, RG4004, https://doi.org/10.1029/2010RG000345.

Haughton, N., G. Abramowitz, A. Pitman, and S. J. Phipps, 2015: Weighting climate model ensembles for mean and variance estimates. *Climate Dyn.*, **45**, 3169–3181, https://doi.org/10.1007/s00382-015-2531-3.

Henley, B. J., J. Gergis, D. J. Karoly, S. Power, J. Kennedy, and C. K. Folland, 2015: A tripole index for the interdecadal Pacific oscillation. *Climate Dyn.*, **45**, 3077–3090, https://doi.org/10.1007/s00382-015-2525-1.

Holmes, C. R., T. Woollings, E. Hawkins, and H. de Vries, 2016: Robust future changes in temperature variability under greenhouse gas forcing and the relationship with thermal advection. *J. Climate*, **29**, 2221–2236, https://doi.org/10.1175/JCLI-D-14-00735.1.

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

Kattsov, V. M., and J. E. Walsh, 2000: Twentieth-century trends of Arctic precipitation from observational data and a climate model simulation. *J. Climate*, **13**, 1362–1370, https://doi.org/10.1175/1520-0442(2000)013<1362:TCTOAP>2.0.CO;2.

——, ——, W. L. Chapman, V. A. Govorkova, T. V. Pavlova, and X. Zhang, 2007: Simulation and projection of Arctic freshwater budget components by the IPCC AR4 global climate models. *J. Hydrometeor.*, **8**, 571–589, https://doi.org/10.1175/JHM575.1.

Knight, J. R., 2009: The Atlantic multidecadal oscillation inferred from the forced climate response in coupled general circulation models. *J. Climate*, **22**, 1610–1625, https://doi.org/10.1175/2008JCLI2628.1.

Kravtsov, S., 2017: Comment on "Comparison of low-frequency internal climate variability in CMIP5 models and observations." *J. Climate*, **30**, 9763–9772, https://doi.org/10.1175/JCLI-D-17-0438.1.

——, and D. Callicutt, 2017: On semi-empirical decomposition of multidecadal climate variability into forced and internally generated components. *Int. J. Climatol.*, **37**, 4417–4433, https://doi.org/10.1002/joc.5096.

——, M. G. Wyatt, J. A. Curry, and A. A. Tsonis, 2015: Comment on "Atlantic and Pacific multidecadal oscillations and Northern Hemisphere temperatures." *Science*, **350**, 1326, https://doi.org/10.1126/science.aab3570.

Maher, N., S. McGregor, M. H. England, and A. Sen Gupta, 2015: Effects of volcanism on tropical variability. *Geophys. Res. Lett.*, **42**, 6024–6033, https://doi.org/10.1002/2015GL064751.

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

——, and K. A. Emanuel, 2006: Atlantic hurricane trends linked to climate change. *Eos, Trans. Amer. Geophys. Union*, **87**, 233–244, https://doi.org/10.1029/2006EO240001.

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

Olonscheck, D., and D. Notz, 2017: Consistently estimating internal climate variability from climate model simulations. *J. Climate*, **30**, 9555–9573, https://doi.org/10.1175/JCLI-D-16-0428.1.

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, https://doi.org/10.1038/ngeo955.

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, https://doi.org/10.1029/2002JD002670.

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

Sillmann, J., V. V. Kharin, F. W. Zwiers, X. Zhang, and D. Bronaugh, 2013: Climate extremes indices in the CMIP5 multimodel ensemble: Part 2. Future climate projections. *J. Geophys. Res. Atmos.*, **118**, 2473–2493, https://doi.org/10.1002/jgrd.50188.

Steinman, B. A., L. M. Frankcombe, M. E. Mann, S. K. Miller, and M. H. England, 2015a: Response to comment on "Atlantic and Pacific multidecadal oscillations and Northern Hemisphere temperatures." *Science*, **350**, 1326, https://doi.org/10.1126/science.aac5208.

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

Swingedouw, D., J. Mignot, P. Ortega, M. Khodri, M. Menegoz, C. Cassou, and V. Hanquiez, 2017: Impact of explosive volcanic eruptions on the main climate variability modes. *Global Planet. Change*, **150**, 24–45, https://doi.org/10.1016/j.gloplacha.2017.01.006.

Taschetto, A. S., A. Sen Gupta, N. C. Jourdain, A. Santoso, C. C. Ummenhofer, and M. H. England, 2014: Cold tongue and warm pool ENSO events in CMIP5: Mean state and future projections. *J. Climate*, **27**, 2861–2885, https://doi.org/10.1175/JCLI-D-13-00437.1.

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, https://doi.org/10.1175/BAMS-D-11-00094.1.

Trenberth, K. E., and D. J. Shea, 2006: Atlantic hurricanes and natural variability in 2005. *Geophys. Res. Lett.*, **33**, L12704, https://doi.org/10.1029/2006GL026894.

Wyatt, M. G., S. Kravtsov, and A. A. Tsonis, 2012: Atlantic multidecadal oscillation and Northern Hemisphere's climate variability. *Climate Dyn.*, **38**, 929–949, https://doi.org/10.1007/s00382-011-1071-8.