Comment on "Comparison of low-frequency internal climate variability in CMIP5 models and observations"

by

Sergey Kravtsov[1]

University of Wisconsin-Milwaukee

28 June 2017

Submitted to the *Journal of Climate*

[1]*Corresponding author address*: Dept. of Mathematical Sciences, Atmospheric Science Group, University of Wisconsin-Milwaukee, P. O. Box 413, Milwaukee, WI 53201. *E-mail*: kravtsov@uwm.edu.

Abstract

In a recent article (*J. Climate*, doi:10.1175/JCLI-D-16-0712.1), Cheung et al. (2017) [hereafter, C2017] apply a semi-empirical methodology to isolate internal climate variability (ICV) in CMIP5 models and observations. The essence of their methodology is to subtract scaled CMIP5 multi-model ensemble-mean (MMEM) from individual model simulations and from the observed time series of several surface-temperature indices. C2017 detect large differences in both the magnitude and spatial patterns of the observed and simulated ICV, as well as large differences between the historical (simulated) ICV and pre-industrial (PI) control CMIP5 simulations. Here it is shown that subtraction of the scaled MMEM from CMIP5 historical simulations produces a poor estimate of the modeled ICV due to the difference between the scaled MMEM and a given model's true forced signal masquerading as ICV. The resulting phase and amplitude errors of the ICV so estimated are large, which compromises most of the C2017 conclusions. By contrast, an alterative methodology based on forced signals computed from individual model ensembles produces a much more accurate estimate of the ICV in CMIP5 models, whose magnitude is consistent with the PI control simulations and is much smaller than any of the semi-empirical estimates of the observed ICV on decadal and longer time scales.

1. Introduction

Accurate identification of internal (unforced) climate variability (hereafter, ICV) is essential for understanding and predicting long-term climate change apparent, for example, in a variety of the Northern Hemisphere surface-temperature indices. One of the ways to isolate ICV in climate-model historical simulations is to utilize ensembles of simulations available through the Coupled Model Intercomparison Project phase 5 (CMIP5: Taylor et al. 2012). Here, averaging historical simulations over multiple realizations of a climate-index time series of a given model provides an estimate of the forced signal, while differencing each simulation with the forced signal so computed provides an estimate of the simulated ICV. Steinman et al. (2015a,b) correctly noted that forced signals inferred via such single-model ensemble means (SMEM) are contaminated by insufficient averaging of ICV — due to a small number of historical realizations typically available in individual model ensembles, — and argued that additional averaging across the entire multi-model ensemble (multi-model ensemble mean: MMEM) is required for proper estimation of the true forced signal. They claimed that the MMEM rescaled, via linear regression, — to match either the observed or simulated climate change — provides unbiased estimates of the forced signal in observations and CMIP5 historical simulations, and can thus be used to obtain unbiased estimates of (both observed and simulated) residual ICV. Along the same lines, Frankcombe et al. (2015) introduced two-factor and three-factor scaling methods for identification of the forced signal and ICV; these methods utilized, in addition to the MMEM based on historical simulations incorporating all forcing factors, the MMEMs from the historical simulations forced by the greenhouse-gas or natural forcing only. In this comment, we will refer to the single-

factor, two-factor and three-factor estimates of the forced signal and the associated estimates of ICV as the MMEM1, MMEM2 and MMEM3 methods, respectively.

Recently, Kravtsov et al. (2015) and Kravtsov and Callicutt (2017) showed that Steinman et al. (2015a,b) arguments are flawed, and it is impossible to characterize a rich spectrum of forced signals realized in the individual CMIP5 models using the single rescaled MMEM signal. In particular, these authors demonstrated that subtraction of the rescaled MMEM from the ensembles of individual-model simulations does not produce independent (uncorrelated) realizations of this model's ICV, which are instead contaminated (and often dominated) by the common difference between this model's true forced signal and the rescaled MMEM. They also showed that smoothed (5-yr low-pass filtered) SMEM of an individual model provides a much more accurate estimate of this model's true forced signal. Kravtsov and Callicutt (2017) further used the appropriately rescaled SMEMs of individual models in conjunction with the observed time series of a given climate index to estimate forced signal (and ICV) in observations. Meanwhile, Kravtsov (2017) developed a spectral variance-inflation method to correct for the errors associated with insufficient SMEM averaging and to produce unbiased estimates of ICV in CMIP5 historical simulations. We will hereafter jointly refer to the above methods for estimating the observed and simulated ICV over the historical period as KC2017.

The above criticisms of the MMEM1 method also fully apply to a recent work by Cheung et al. (2017) [hereafter C2017], who used it as the main tool to compare the observed and CMIP5 simulated ICV of surface temperature over the course of the twentieth century. These authors found considerable mismatches in the magnitude of the observed and modeled ICV on decadal and longer time scales, as well as

mismatches in the ICV between CMIP5 historical runs and pre-industrial (PI) control runs, with the simulated ICV over the historical period being generally much lower than the observed ICV, but higher than the variability in the PI control runs. The main point of this comment is to show that these results are quantitatively inaccurate due to biases of the MMEM1 method, whereas KC2017 method provides means to a much more accurate estimation of both the amplitude and the phase of the true ICV in historical simulations.

Further presentation is organized as follows. The data and analysis methodology are described in section 2. Section 3 analyzes the performance of the various MMEM methods, as well as KC2017 method in isolating ICV in the simulations from the Community Earth System Model (CESM) Large Ensemble Project (LENS: Kay et al. 2015): http://www.cesm.ucar.edu/projects/ community-projects/LENS/, in which a large ensemble size permits a fairly accurate determination of the true forced (and internal) variability. The application of these methods to CMIP5 multi-model ensemble is discussed in section 4, and conclusions are presented in section 5. Supplemental Materials contain further information in support of this paper's main results, as well as the link to the data and code for all of the analyses performed here.

2. Data and methods

We utilized the same historical (see **Table 1**), historicalNat (natural forcing only), historicalGHG (greenhouse-gas forcing only) and PI control CMIP5 simulations (1880–2005) as in Kravtsov (2017). The sub-ensemble of CMIP5 historical simulations (17 models, 111 simulations) was smaller than the ensemble

used in C2017 since we only considered the models with four or more historical simulations available to be able to apply the KC2017 method; note, however, that the MMEM based on this sub-ensemble is virtually indistinguishable from the MMEM based on the entire CMIP5 ensemble (not shown). We also utilized the 40-member historical ensemble (1920–2005) and long PI control simulations from the LENS project (Kay et al. 2015). The observed surface-temperature data we used was also identical to the data used in Kravtsov (2017) [and C2017]; one notable difference with C2017 here is our use of the twentieth-century reanalysis (20CR; Compo et al. 2011) to compute the Northern Hemisphere mean temperature (instead of GISTEMP data in C2017). Our results do not depend on the choice of surface-temperature data set.

Following Steinman et al. (2015a) and C2017, we considered Atlantic Multidecadal Oscillation (AMO), Pacific Multidecadal Oscillation (PMO) and Northern Hemisphere Multidecadal Oscillation (NMO) indices; the AMO and PMO indices were computed as sea-surface temperature averaged over the North Atlantic and North Pacific (0º–60ºN), respectively, and the NMO was computed as a weighted sum of the 0º–60ºN surface air temperature averaged over ocean and land, with the weights set to 0.61 and 0.39, respectively.

We computed forced signals and ICV estimates (in both observations and model simulations) using MMEM1, MMEM2, MMEM3 and KC2017 methods. In applying KC2017 method to LENS simulations, we considered random sub-ensembles of five simulations (which is a typical size of an individual model ensemble in CMIP5 models) containing a given simulation to compute 5-yr low-pass filtered ensemble mean as a first-guess estimate of the forced signal and the residual ICV in that simulation, then applied this procedure to all other simulations. The resulting 40 estimates of ICV were further variance-bias corrected using the

procedure developed in Kravtsov (2017). For the LENS ensemble, we also computed the "true" forced signal as the grand (40-member) ensemble mean, and the associated "true" residual ICV for each of the 40 historical realizations. As in C2017, we further computed low-pass filtered versions of the AMO, PMO and NMO ICV using the data-adaptive filter of Mann (2008) with cutoff frequencies of 0.1, 0.05 and 0.025 cycles per year (or corresponding smoothing time scales of 10, 20 and 40 years).

3. Quantifying performance of different methods for isolating internal variability in historical simulations using CESM LENS ensemble

Figure 1 shows the magnitude of internal variability in LENS simulations estimated by different methods (top row), as well as the root-mean-square (rms) error of each method with respect to the "true" internal variability based on the subtraction of the grand ensemble mean from individual simulations (bottom row), as a function of smoothing time scale. It turns out that the rescaled CMIP5 MMEM is fairly close to the true forced signal in the CESM model, and the MMEM-based methods generally provide decent estimates of ICV in CESM historical runs (see **Fig. S1** of Supplemental Materials). Still, the KC2017 method is more accurate than MMEM1 method in identifying low-frequency ICV, with the rms error of the 40-yr low-pass filtered ICV estimated using KC2017 method being 75% of the MMEM1 error for PMO (Fig. 1b), and about 40% of the MMEM1 error for AMO and NMO (Figs. 1d, f). The MMEM2 and MMEM3 methods perform worse than both MMEM1 and KC2017 methods across the entire range of time scales and for all indices. The MMEM1 and MMEM2 methods tend to overestimate the magnitude of the true ICV at low frequencies, whereas the MMEM3 method slightly underestimates this magnitude

(Figs. 1a,c,e). Note that the amplitudes of ICV estimated by all methods are overall consistent; hence, large rms errors of the MMEM2 and MMEM3 methods (Figs. 1b,d,f) are due to a combination of both amplitude and, most importantly, phase errors with respect to the true ICV (in other words, the magnitude of the estimated ICV may match the magnitude of the true ICV, but it may still be poorly correlated with the true ICV; see examples in Fig. S1 of Supplemental Materials). In summary, KC2017 method provides the best estimate of the ICV in CESM LENS simulations, whereas the performance of MMEM-based methods strongly depends on how close a given type of MMEM is to the true forced signal of this model.

4. Internal variability in observations and CMIP5 simulations: Comparison of different estimation methods

The main problem with using MMEM-based methods for estimation of ICV in CMIP5 models, as in C2017, is illustrated in **Fig. 2**, which shows estimates of ICV in the five available historical simulations of GFDL CM3 model. Based on the results of section 3, we assume that the KC2017 method provides the best estimate of the ICV; Kravtsov et al. (2015) showed that this method indeed leads to the independent (uncorrelated) ICV estimates in different realizations of a given model. By contrast, all of the GFDL CM3's realizations of ICV estimated by using MMEM1 and MMEM2 methods clearly have the same shape (and are thus strongly correlated) in the second half of the twentieth century, where the true forced signal of the GFDL CM3 model is very different from its MMEM-based estimate (not shown). The internal variability estimated using these methods is thus contaminated (and in this case dominated) by the mismatch between the true forced signal of GFDL CM3 model and the MMEM. The MMEM3 method does much better, and its estimated

ICV is closer to the KC2017 estimates (because the MMEM3 based forced signal is closer to that estimated using KC2017 method), but there are still considerable differences in the phase and amplitude of ICV estimated by the MMEM3 and KC2017 methods.

Figure 3 summarizes the performance of the four methods discussed in this paper across the entirety of the CMIP5 sub-ensemble considered. We concentrate here on the results for the 40-yr low-pass filtered ICV, as the results utilizing 10-yr and 20-yr smoothing time scale are qualitatively similar. The results for the PMO, AMO and NMO indices are also consistent. First off, irrespective of the method used to infer either observed or simulated ICV, the magnitude of multidecadal ICV in CMIP5 simulations is much lower than the observed magnitude, consistent with Frankcombe et al. (2015), Kravtsov and Callicutt (2017), Kravtsov (2017) and C2017. Among the observed estimates, KC2017 method leads to the largest estimated magnitude of the ICV, and MMEM3 method — to the smallest magnitude, while the magnitude estimated by the MMEM1 and MMEM2 methods is in between. For the simulated variability, the magnitude of the ICV inferred via MMEM1 and MMEM2 methods is the largest, and exceeds the magnitude of the variability in PI control runs by a factor of ~1.7–1.8 for PMO and NMO indices, and by a factor of 1.25 for AMO index, consistent with C2017. Both KC2017 and MMEM3 methods give magnitudes of simulated ICV consistent with the PI control runs, with KC2017-based magnitude being slightly larger, and MMEM3-based magnitude — slightly smaller than the magnitude in the PI control simulations.

The amplitude bias in the MMEM1 and MMEM2 based estimates of ICV in historical CMIP5 simulations stems from misidentification of the ICV illustrated in

Fig. 2, as the individual models' forced signals inferred using these methods are systematically different from the true forced signals in these models, which are best approximated by the KC2017 method. Of the three MMEM-based methods considered, the MMEM3 method provides the estimates of the forced signal and ICV that best match the KC2017 estimates (Figs. 3b,d,f), but even in this case, the mismatch between these two methods on multidecadal time scales (for the 40-yr low-pass filtered data) is characterized by a large rms error on the order of the standard deviation of the simulated ICV (compare brown bars in Figs. 3b,d,f with the green or brown bars in Figs. 3a,c,e, respectively).

In general, the MMEM-based methods are expected to provide faithful estimates of the true ICV in a given model if this model's true forced signal — well approximated by the KC2017 method — is close to the corresponding scaled MMEMs, and poor estimates otherwise. We thus developed a similarity index to characterize the individual models in terms of how well their MMEM-based estimates of the forced signal and ICV match the corresponding KC2017 estimates. To do so, we utilized 40-yr low-pass filtered MMEM1, MMEM2, MMEM3 and KC2017 estimates of ICV and computed the average rms error of the three MMEM-based methods with respect to the KC2017 method for individual model sub-ensembles. For a given climate index (AMO, PMO or NMO) and a given MMEM method (MMEM1, MMEM2 or MMEM3), this procedure resulted in 17 values of the rms error (one value per model). We then assigned the consistency index value of +1 to the models characterized by the lowest rms (below the 33rd percentile of the 17 rms values), the consistency index value of –1 to the models characterized by the highest rms (above the 66th percentile of the 17 rms values) and the index value of 0 to all other models. Finally, we averaged the resulting consistency index values across the estimates

corresponding to the MMEM1, MMEM2 and MMEM3 methods, and then across the consistency index estimates for AMO, PMO and NMO time series. The resulting consistency index has a range between –1 and +1. The value of –1 corresponds to the situation when a given model is characterized by a high rms error with respect to the KC2017 method for all MMEM-based methods and for all three climate indices (hence, for such a model, the MMEM-based estimates of the forced signal and internal variability are poor). Conversely, the consistency index value of +1 would indicate that the corresponding model's true forced signal is close to its MMEM-based estimate. We further clustered the models in three groups by selecting the models with consistency indices below –0.34, above +0.34, and in between, and assigning to these models the values of –1 (true forced signal inconsistent with MMEM), +1 (true forced signal consistent with MMEM), and 0 (intermediate group in terms of consistency between MMEM-based and KC2017-based forced signals).

The consistency index results (Table 1) indicate that MMEM-based forced signal estimation methods produce good estimates of the ICV in GISS models (except for Rp3), as well as in MRI-CGCM3, poor estimates of ICV in CanESM2, CSIROmk360, GFDL-CM3, GISSE2Rp3, HadGEM2-ES and IPSL-CM5A-LR models, and intermediate-quality estimates in CCSM4, CNRMCM5, GFDL-CM2p1, HadCM3 and MIROC5 models. Figure 2 gives an example of the ICV in a model from the 'poor' group; see **Figs. S2 and S3** of Supplemental Materials for the analogous examples from 'good' and 'intermediate' groups. Thus, CMIP5 models are characterized by a wide variety of forced signals, which MMEM-based estimation methods fail to capture (compare with Kravtsov et al. 2015; Kravtsov and Callicutt 2017; Kravtsov 2017). Therefore, application of these methods (and especially MMEM1 method) to characterize ICV in historical CMIP5 simulations — as was

done in C2017 — is not appropriate.

5. Conclusions

We compared the performance of four methods for estimating internal climate variability (ICV) in historical simulations of CMIP5 climate models. The methods considered included a suite of methods based on subtraction of scaled multi-model ensemble means (MMEM) from individual model simulations or observations (MMEM1, MMEM2, MMEM3 methods; Steinman et al. 2015a,b, Frankcombe et al. 2017; Cheung et al. 2017 [C2017]), as well as the KC2017 method (Kravtsov et al. 2015; Kravtsov and Callicutt 2017; Kravtsov 2017), which works with smoothed ensemble means of individual models. We found that:

1. KC2017 method provides the best estimates of the low-frequency (10-yr+) ICV in historical simulations of individual models, and results in independent (uncorrelated) time series of ICV in individual simulations of a given model. The magnitude of ICV so computed is consistent with that of the variability in the PI control runs.
2. The performance of MMEM-based methods depends on whether the scaled MMEM signal is close to this model's true forced signal or not. In the latter case, the estimated ICV is contaminated, and often dominated by the difference between the true forced signal and its MMEM-based estimate, resulting in large phase and amplitude errors of the ICV so inferred. The resulting estimates of the ICV in individual simulations of a given model are not independent (correlated). For MMEM1 method, the magnitude of the estimated ICV is larger than that in the PI control runs.

3. Out of the 17 CMIP5 models considered here, only 6 models are characterized by the forced signals well represented by the MMEM methods. For the rest of the models, the differences between the scaled MMEMs and their true forced signal estimates obtained by the KC2017 method are large, and the resulting MMEM-based estimates of the ICV are strongly biased as per item 2 above. Hence, the ICV in the ensemble of historical CMIP5 simulations cannot be reliably determined using MMEM-based methods.
4. Out of the three MMEM-based methods applied to the CMIP5 ensemble, the three-factor scaling method (MMEM3) performs, on average, better than one-factor and two-factor methods (MMEM1 and MMEM2), but is still characterized by large errors relative to the KC2017 method.
5. All methods indicate that the estimated observed ICV at decadal and longer time scales has a much larger magnitude than the CMIP5 simulated ICV. The disparity between the observed and simulated ICV is most pronounced in the KC2017 estimates of this variability.

Our results demonstrate that application of the MMEM1 method to infer ICV in historical CMIP5 simulations in C2017 work is inadequate, and leads to large phase and amplitude errors of ICV so estimated, which invalidates most of these authors' conclusions.

Acknowledgements. We acknowledge the World Climate Research Programme's Working Group on Coupled Modeling, which is responsible for CMIP, and we thank the climate modeling groups for producing and making available their model output. This research was supported by the NSF grant AGS-1408897. All data and MATLAB © scripts for this paper are available for downloading using the link listed in the Supplemental Materials.

References

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

Compo, G. P., et al., 2011: The twentieth century reanalysis project. *Quart. J. Royal Meteor. Soc.*, **654**, 1–28, doi: 10.1002/qj.776.

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, doi: http://dx.doi.org/10.1175/JCLI-D-15-0069.1.

Kay, J. E., et al., 2015: The Community Earth System Model (CESM) Large Ensemble Project: A community resource for studying climate change in the presence of internal climate variability. *Bull. Amer. Meteor. Soc.*, **96**, 1333–1349.

Kravtsov, S., 2017: Pronounced differences between observed and CMIP5-simulated multidecadal climate variability in the twentieth century. *Geophys. Res. Lett.*, published online 15 June 2017, doi: 10.1002/2017GL074016, http://onlinelibrary.wiley.com/doi/10.1002/2017GL074016/full.

Kravtsov, S., and D. Callicutt, 2017: On semi-empirical decomposition of multidecadal climate variability into forced and internally generated components. *Internat. J. Climatol.*, published online April 2017, doi: 10.1002/joc.5096, http://onlinelibrary.wiley.com/doi/10.1002/joc.5096/full.

Kravtsov, S., 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, doi: 10.1126/science.aab3570.

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

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

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

Taylor, K. E., R. J. Stouffer and G. A. Meehl, 2012: An Overview of CMIP5 and the experiment design. *Bull. Am. Meteorol.* Soc., **93,** 485–498.

Table captions

Table 1. Classification of the CMIP5 models considered in terms of consistency between their internal variability estimates computed using MMEM-based methods and KC2017 method. The first two columns give model acronyms and the number of historical simulations for a given model. The consistency indices themselves — defined to range from –1 (low consistency between the two methods) to +1 (high consistency between the two methods) — are listed in the last two columns, with the first of these columns showing the raw index, and the second — its ternary version. See text for details.

Figure captions

Figure 1: Performance of different methods for isolating internal variability in 40 historical simulations (1920–2005) of the Community Earth System Model (CESM) Large Ensemble Project (LENS). The methods are one-factor (MMEM1), two-factor (MMEM2) and three-factor (MMEM3) methods based on subtraction of scaled CMIP5 multi-model ensemble-mean (MMEM) forced-signal estimates from individual LENS simulations, as well as KC2017 method, applied here to random 5-member sub-ensembles of LENS simulations. Top row: standard deviation of the raw and low-pass-filtered estimates of internal variability, as a function of smoothing time scale. In addition to estimates for all methods above, also included here are the estimates for the "true" internal variability based on subtracting LENS grand ensemble mean from individual simulations, as well as estimates based on random 86-yr-long snippets from CESM pre-industrial (PI) control run (see panel legends). Bottom row: root-mean-square (rms) errors of internal variability estimated using the four methods with respect to the "true" internal variability (see panel legends), again as a function of smoothing time scale. Bars show the mean value and error bars — the standard deviation of each quantity across the 40-member LENS ensemble. (a, b) Results for PMO index; (c, d) results for AMO index; (e, f) results for NMO index.

Figure 2: Estimates of internal variability in historical simulations of the GFDL CM3 model using (from top to bottom) MMEM1, MMEM2, MMEM3, KC2017 methods; also shown are five random samples of internal variability from the PI control run (bottom row). All time series were 40-yr low-pass filtered using

data-adaptive filter of Mann (2008) and windowed using tapers from Kravtsov and Callicutt (2017) to reduce spurious end effects. Left column: results for AMO; middle column: results for PMO; right column: results for NMO.

Figure 3: Internal variability in observations and CMIP5 simulations. Top row: magnitude (standard deviation) of the observed and simulated internal variability estimated using four different methods, as well as estimates based on random 126-yr-long snippets from PI control runs (see panel legends), as a function of smoothing time scale. Bottom row: rms errors of internal variability estimated using various MMEM methods with respect to the KC2017 estimate of the internal variability (see panel legends), again as a function of smoothing time scale. Bars show the mean value and error bars — the standard deviation of each quantity across the entire ensemble of simulations. (a, b) Results for PMO index; (c, d) results for AMO index; (e, f) results for NMO index.

Table 1. Classification of the CMIP5 models considered in terms of consistency between their internal variability estimates computed using MMEM-based methods and KC2017 method. The first two columns give model acronyms and the number of historical simulations for a given model. The consistency indices themselves — defined to range from –1 (low consistency between the two methods) to +1 (high consistency between the two methods) — are listed in the last two columns, with the first of these columns showing the raw index, and the second — its ternary version. See text for details.

Model acronym	Number of realizations	Consistency between MMEM and KC17 methods	
CanESM2	5	–0.44	–1
CCSM4	6	0	0
CNRM-CM5	10	0.22	0
CSIRO-Mk3-6-0	10	–1	–1
GFDL-CM2p1	10	–0.33	0
GFDL-CM3	5	–0.78	–1
GISS-E2-Hp1	6	0.67	1
GISS-E2-Hp2	6	0.89	1
GISS-E2-Hp3	6	0.89	1
GISS-E2-Rp1	6	1	1
GISS-E2-Rp2	6	0.67	1
GISS-E2-Rp3	6	–0.44	–1
HadCM3	10	–0.33	0
HadGEM2-ES	5	–0.55	–1
IPSL-CM5A-LR	6	–0.78	–1
MIROC5	5	–0.11	0
MRI-CGCM3	3	0.44	1

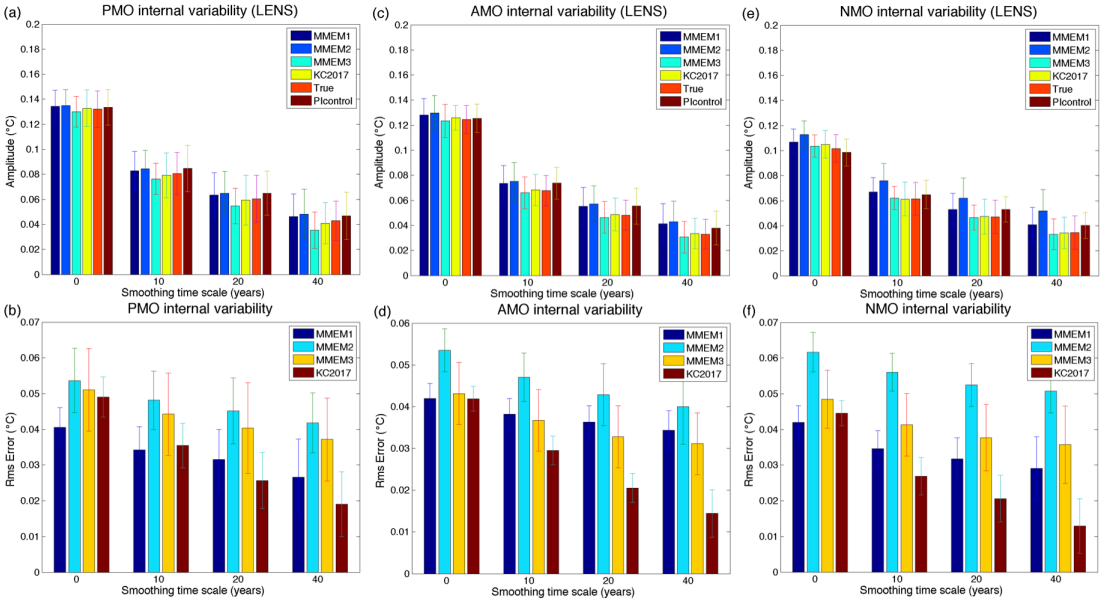

Figure 1: Performance of different methods for isolating internal variability in 40 historical simulations (1920–2005) of the Community Earth System Model (CESM) Large Ensemble Project (LENS). The methods are one-factor (MMEM1), two-factor (MMEM2) and three-factor (MMEM3) methods based on subtraction of scaled CMIP5 multi-model ensemble-mean forced-signal estimates from individual LENS simulations, as well as KC2017 method, applied here to random 5-member sub-ensembles of LENS simulations. Top row: standard deviation of the raw and low-pass-filtered estimates of internal variability, as a function of smoothing time scale. In addition to estimates for all methods above, also included here are the estimates for the "true" internal variability based on subtracting LENS grand ensemble mean from individual simulations, as well as estimates based on random 86-yr-long snippets from CESM pre-industrial (PI) control run (see panel legends). Bottom row: root-mean-square (rms) errors of internal variability estimated using the four methods with respect to the "true" internal variability (see panel legends), again as a function of smoothing time scale. Bars show the mean value and error bars — the standard deviation of each quantity across the 40-member LENS ensemble. (a, b) Results for PMO index; (c, d) results for AMO index; (e, f) results for NMO index.

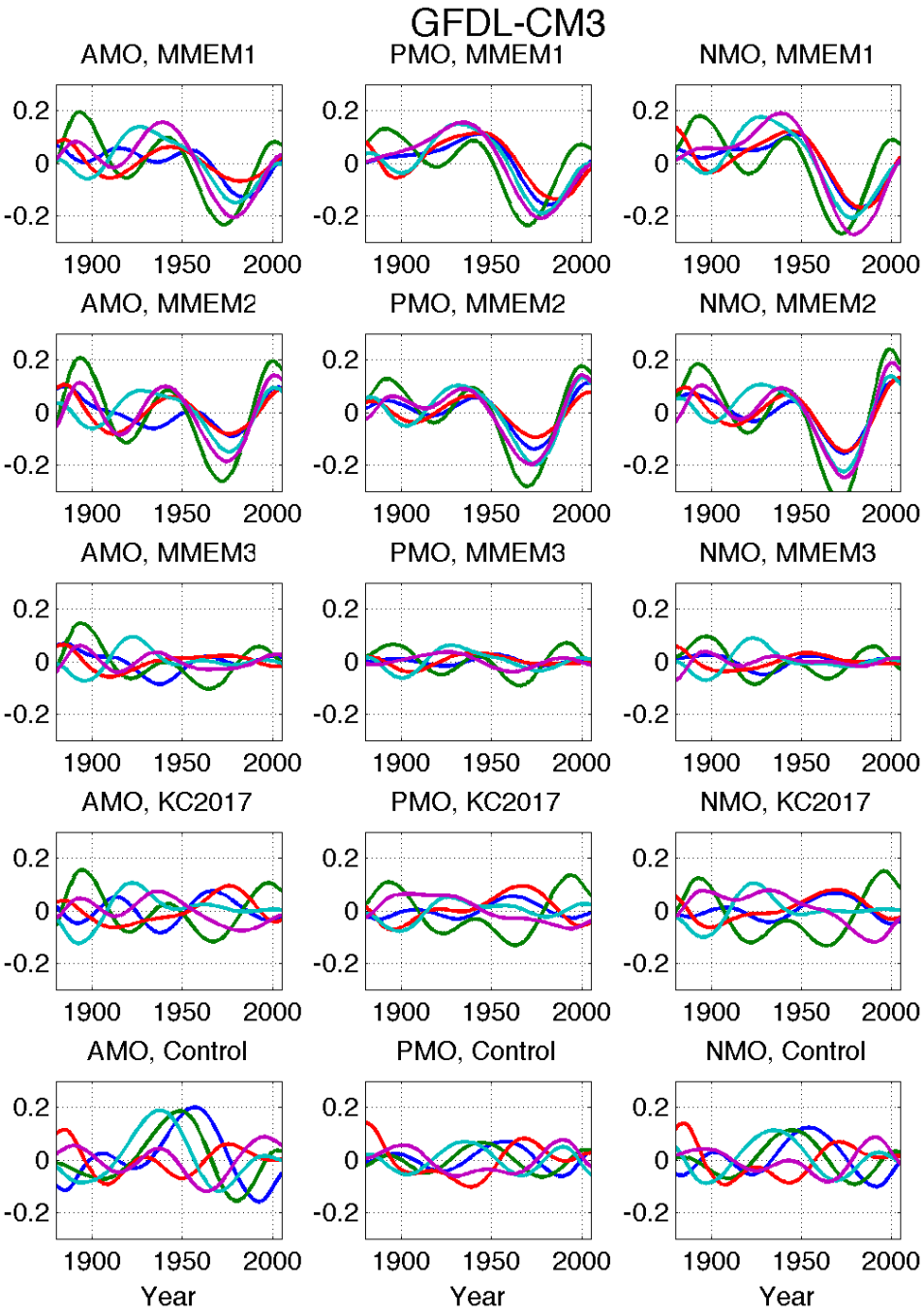

Figure 2: Estimates of internal variability in historical simulations of the GFDL CM3 model using (from top to bottom) MMEM1, MMEM2, MMEM3, KC2017 methods; also shown are five random samples of internal variability from the PI control run (bottom row). All time series were 40-yr low-pass filtered using data-adaptive filter of Mann (2008) and windowed using tapers from Kravtsov and Callicutt (2017) to reduce spurious end effects. Left column: results for AMO; middle column: results for PMO; right column: results for NMO.

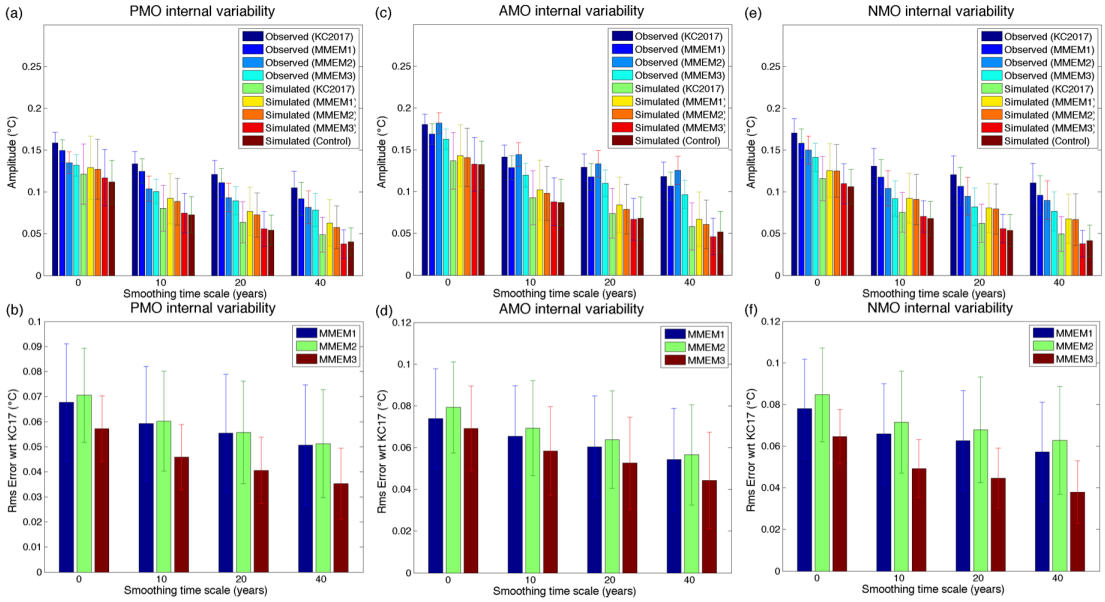

Figure 3: Internal variability in observations and CMIP5 simulations. Top row: magnitude (standard deviation) of the observed and simulated internal variability estimated using four different methods, as well as estimates based on random 126-yr-long snippets from PI control runs (see panel legends), as a function of smoothing time scale. Bottom row: rms errors of internal variability estimated using various MMEM methods with respect to the KC2017 estimate of the internal variability (see panel legends), again as a function of smoothing time scale. Bars show the mean value and error bars — the standard deviation of each quantity across the entire ensemble of simulations. (a, b) Results for PMO index; (c, d) results for AMO index; (e, f) results for NMO index.