TECHNICAL COMMENT

CLIMATE CHANGE

Comment on "Atlantic and Pacific multidecadal oscillations and Northern Hemisphere temperatures"

S. Kravtsov,[1]* M. G. Wyatt,[2] J. A. Curry,[3] A. A. Tsonis[1]

Steinman *et al*. (Reports, 27 February 2015, p. 988) argue that appropriately rescaled multimodel ensemble-mean time series provide an unbiased estimate of the forced climate response in individual model simulations. However, their procedure for demonstrating the validity of this assertion is flawed, and the residual intrinsic variability so defined is in fact dominated by the actual forced response of individual models.

The central result of Steinman *et al*.'s analysis (*1*) is the demonstration of an apparent consistency among the responses of different models to variable forcing in the 20th-century climate simulations. In particular, they claim that regional multimodel ensemble-mean time series defines the universal forced signal, which can be linearly rescaled to provide unbiased estimates of the regional forced responses for individual models. Such a consistency is surprising because the models have different physical parameterizations and the simulations may use different forcing subsets. If their claim were true, it would add much confidence to the authors' semi-empirical attribution of the observed multidecadal climate variability to the forced and intrinsic sources. However, the implied uniqueness of the forced signal defined by their regional regression method is an artifact of their analysis procedure, and the actual uncertainty of the semi-empirical estimates of the observed multidecadal intrinsic variability is much larger than these authors have inferred.

Consider M time series of length T, corresponding to M different climate simulations: $x_m^{(t)}; m = 1, ..., M; t = 1, ..., T$. Let the bar denote averaging across the time dimension (t) and square brackets denote averaging across the ensemble member dimension (m). For example, the time mean of each ensemble member $\overline{x_m}$ and the ensemble average time series $[x^{(t)}]$ are defined as follows

$$\overline{x_m} = \frac{1}{T}\sum_{t=1}^{T} x_m^{(t)} \qquad (1)$$

$$[x^{(t)}] = \frac{1}{M}\sum_{m=1}^{M} x_m^{(t)} \qquad (2)$$

Consider a decomposition of $x_m^{(t)}$ into the forced signal $f_m^{(t)}$ and residual intrinsic variability $\epsilon_m^{(t)}$

$$x_m^{(t)} = f_m^{(t)} + \epsilon_m^{(t)} \qquad (3)$$

Without loss of generality, we can assume $\overline{x_m} = \overline{f_m} = 0$, hence $\overline{\epsilon_m} = 0$. If the estimated forced signal $f_m^{(t)}$ is unbiased, then the time series $\epsilon_{m_1}^{(t)}$ and $\epsilon_{m_2}^{(t)}$ of residual intrinsic variability in any pair of simulations m_1 and m_2 must be uncorrelated (independent). Furthermore, if the distribution of $\epsilon_m^{(t)}$ has mean 0 and variance σ^2, the ensemble mean residual time series $[\epsilon^{(t)}]$ will have the distribution with mean 0 and variance σ^2/M. Hence, one can quantitatively assess the statistical independence of different realizations of simulated intrinsic variability by comparing the actual dispersion $[\epsilon]^2$ of the ensemble mean time series $[\epsilon^{(t)}]$ with its theoretical prediction $[\overline{\epsilon^2}]/M$, where we estimated $\sigma^2 \sim [\overline{\epsilon^2}]$. Large values of $[\epsilon]^2$ would indicate that assumption of statistical independence between different realizations of intrinsic variability $\epsilon_m^{(t)}$ is violated due to biases in the estimated forced signal $f_m^{(t)}$, so that at least a portion of the common true forced signal manifests in the estimated "intrinsic" residuals $\epsilon_m^{(t)}$.

Steinman *et al*. considered, among others, the following two methods for estimating the forced signal, both based on the multimodel ensemble mean time series

$$f_m^{(t)} = [x^{(t)}] \qquad (4A)$$

$$f_m^{(t)} = a_m [x^{(t)}] \qquad (4B)$$

The differencing method (Eqs. 3 and 4A) simply identifies the forced signal with the multimodel ensemble mean $[x^{(t)}]$. The regression method (Eqs. 3 and 4B) rescales the first-guess forced signal $[x^{(t)}]$ for a given simulation by finding a_m via least squares to minimize $\overline{\epsilon_m^2}$ in Eq. 3.

Steinman *et al*. further claimed that both of these methods provided independent realizations of residual intrinsic variability in climate-model simulations, based on the fact that the resulting variance $[\epsilon]^2$ of the ensemble mean residual time series was much smaller than the theoretical value of $[\overline{\epsilon^2}]/M$. However, it is easy to show that, due to the choice of forcing derived using either Eq. 4A or Eq. 4B, this ensemble mean residual time series is identically zero

$$[\epsilon^{(t)}] = 0; t = 1, ..., T \qquad (5)$$

and so is its variance $[\overline{\epsilon}]^2 = 0$. Hence, the extreme smallness of the dispersion of ensemble average intrinsic variability attributed in (*1*) to the statistical independence of its different realizations is actually an artifact of the algebraic constraint (Eq. 5) [see (*2*–*5*)]. This flaw does not mean that the residuals are necessarily correlated (not independent), but a different test is required to determine that.

We now show directly that the regional regression approach (*1*) of defining the forced signal leads to the correlated samples of residual intrinsic variability in the individual-model ensembles (subensembles of simulations using a single model with fixed physics package and an identical forcing history). For these subensembles, it is the expression (Eq. 4A) that naturally gives an unbiased estimate of the forced variability. We considered 18 such subensembles from the Coupled Model Intercomparison Project Phase 5 (CMIP5) models with four or more 20th-century simulations (*6*), totaling 116 individual simulations out of the 170 available simulations. The multimodel ensemble mean based on these 116 simulations is nearly identical to the one computed using all of the available 170 simulations. We defined two alternative sets of the model-simulated intrinsic variability. In method A, we formed realizations of intrinsic variability by subtracting the 5-year low-pass-filtered ensemble mean of each model from this model's individual simulations (i.e., Eq. 4A applied separately to individual model ensembles). The second set (method B) defined the residual intrinsic variability using the forced signal estimated from regional multimodel regression (*1*) (i.e., Eq. 4B applied to the whole ensemble of 116 simulations).

To quantify independence of different realizations of intrinsic variability in the individual-model ensembles, we introduced an ensemble correlation measure C by summing positive correlations among all possible pairs of an individual model's M ensemble members

$$C = \frac{2}{M(M-1)}\sum_{m>l} C_{ml}\, H(C_{ml}) \qquad (6)$$

where $H(x)$ is the Heaviside step function (*7*); the quantity C ranges from 0 (no positive correlations between individual ensemble members) to 1 (all ensemble members are perfectly correlated). The correlation measure (Eq. 6) was computed for raw and low-pass-filtered intrinsic variability defined using methods A and B [Fig. 1, A to C shows results for the Geophysical

[1]Department of Mathematical Sciences, Atmospheric Science group, University of Wisconsin-Milwaukee, Post Office Box 413, Milwaukee, WI 53201, USA. [2]Department of Geological Sciences, University of Colorado, Boulder, CO, USA. [3]School of Earth and Atmospheric Sciences, Georgia Institute of Technology, Atlanta, GA, USA.
*Corresponding author: kravtsov@uwm.edu

Fluid Dynamics Laboratory (GFDL) CM3 model; see (*8*)]. Method A produces intrinsic variability with C values well within the range expected from random uncorrelated red-noise samples generated using an autoregressive model of order 3 (AR-3) (*9*). In contrast, Steinman *et al.*'s method B results in samples that are significantly correlated due to their systematic difference from the true forced signal.

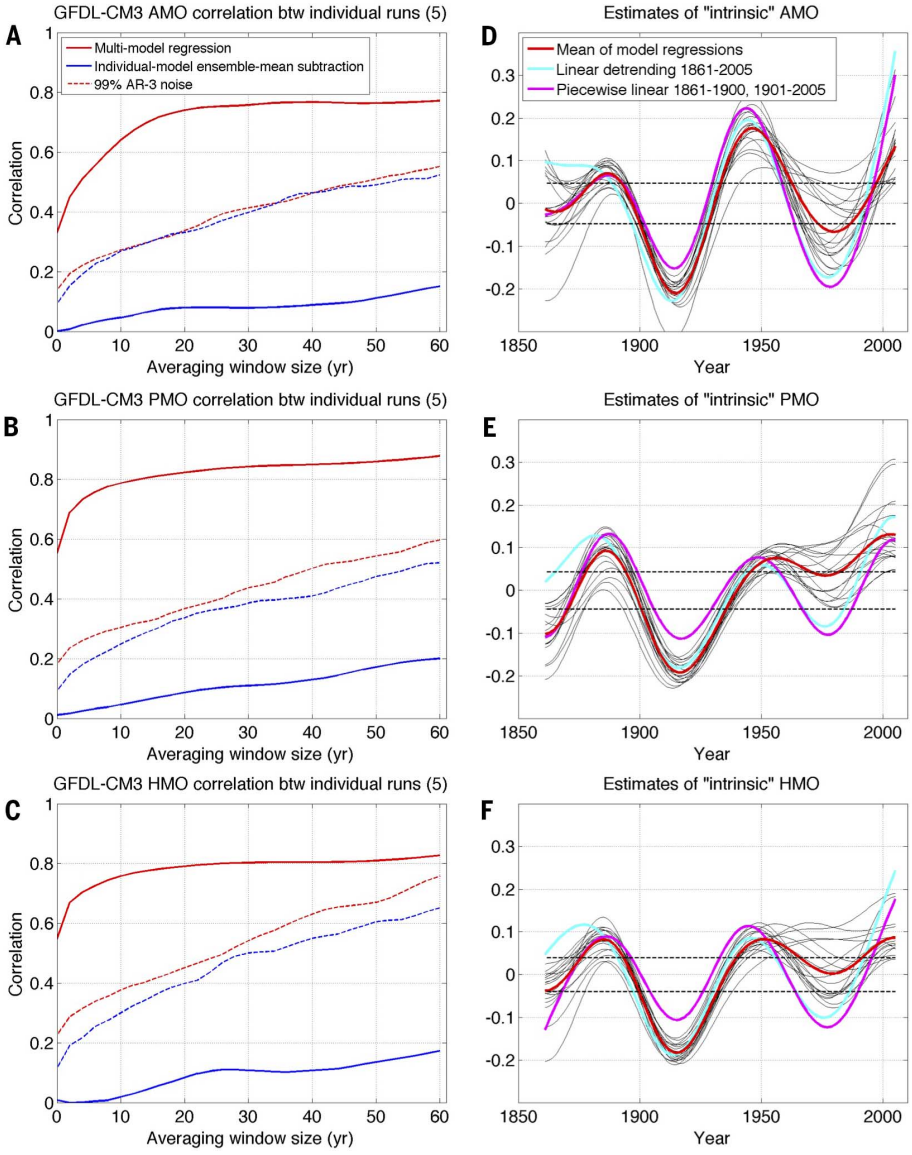

Fig. 1. Intrinsic variability in the 20th-century model simulations with four or more ensemble members identified using two different methods for estimating the forced signal: the classical subtraction of the individual-model ensemble mean (method A) and the multimodel regional regression method (*1*) (method B). (**A** to **C**) The correlation measure (Eq. 6) of statistical independence between multiple realizations of the GFDL CM3 model (five realizations) for (A) Atlantic Multidecadal Oscillation (AMO), (B) Pacific Multidecadal Oscillation (PMO), and (C) Northern Hemisphere Multidecadal Oscillation (HMO) indices; these correlations were computed for running-mean low-pass-filtered residual time series (which characterize intrinsic variability) and are plotted here against the averaging window size. Low correlation measure indicates statistical independence of intrinsic residuals. Dashed lines show the 99th percentile of the correlation measure based on the 1000 simulations of the corresponding AR-3 red-noise model. (**D** to **F**) Estimates of the observed multidecadal intrinsic variability for (D) AMO, (E) PMO, and (F) HMO. The semi-empirical estimates (thin black lines) were computed as in (*1*) based on the forced signals obtained using method A for each of the 18 model ensembles considered, with the heavy red line indicating the average over these individual estimates. Additional heavy lines (see legend) are for results based on linear detrending. The distance between the black dashed lines in each plot shows the 95th percentile of the standard deviations for multidecadal intrinsic variability estimated using method A for each of 116 simulations considered.

We then used 18 versions of the forced signal, estimated by the unbiased method A, to isolate intrinsic variability in observed surface temperatures via Eq. 3 and Eq. 4B (Fig. 1, D to F). The spread among the 18 estimates of intrinsic variability in observations is much larger than the tight bootstrap-based error bounds on the semi-empirical estimates of the observed intrinsic variability in figure 3 in (*1*). Hence, the actual uncertainty of the semi-empirical attribution by SMM is also much larger (*10*), thereby preventing any clear inferences about the cause of the "false pause" in the global warming (*11*, *12*).

REFERENCES AND NOTES

1. B. A. Steinman, M. E. Mann, S. K. Miller, *Science* **347**, 988–991 (2015).
2. The standard deviation of intrinsic variability computed in Steinman *et al.* (*1*) is small, but not exactly zero because of their using a data adaptive low-pass filter before averaging intrinsic variability among different simulations.
3. Steinman *et al.* also used weighted ensemble means to define a version of their model-based forced signal. In this case, the constraint (Eq. 5) would not be exact but would still be approximately valid, because the weighted and nonweighted estimates of the forced signal are in fact very close (not shown here).
4. Comment (*3*) above also applies to a possible variation of the regression method (Eq. 4B) in which, instead of scaling each individual simulation by its own factor a_m, one would estimate and use the single scaling factor for all simulations of each model; this scaling factor can be defined, for example, as the ensemble mean of a_m estimates computed for individual simulations of a given model.
5. One way to try to alleviate constraint (Eq. 5) would be to estimate the forced signal for a given subset of models using the ensemble mean time series of the complement subset of models. However, this would only be effective if the sizes of these two subsets are comparable. Otherwise, the multimodel averaging over the much larger complement subset of models would also be very close to the all-model ensemble mean, and the algebraic constraint (Eq. 5) would still approximately hold.
6. There are 13 models with four or more 20th-century simulations in the CMIP5 data set, but considering separately the ensembles of the Goddard Institute for Space Studies (GISS) models with different physics packages makes up 18 independent ensembles.
7. The Heaviside step function is used here merely to streamline the mathematical notations in the multiple correlation measure (Eq. 6) by zeroing out negative terms in the sum of correlations, leaving positive terms unchanged.
8. Other models exhibit a similar behavior; see the corresponding images at https://pantherfile.uwm.edu/kravtsov/www/downloads/KWCT2015/TIFF_FILES.
9. If one does not divide the large ensembles of the GISS models into the subensembles with different physics (*6*), the correlation-measure diagnosis does identify the dependency between the model realizations, because the true forced responses in these versions of the model are different from the grand ensemble mean response, and similar long-term biases across the same-physics model simulations ensue.
10. The bootstrap resampling used in (*1*) is equivalent to considering subensembles of about two-thirds of independent models (or simulations), thus effectively averaging out the intramodel uncertainty of the forced response emphasized in our Fig. 1, D to F.
11. This is exacerbated further by the unfortunate linear extrapolation of the CMIP5 runs from 2005 to 2012 used in (*1*) to estimate recent intrinsic trends.
12. B. Rajaratnam, J. Romano, M. Tsiang, N. S. Diffenbaugh, *Clim. Change* **133**, 129–140 (2015).

ACKNOWLEDGMENTS

We thank Steinman *et al.* for making their data and analysis code publicly available. This research was supported by NSF grants OCE-1243158 (S.K.) and AGS-1408897 (S.K. and A.A.T.). All data and MATLAB (MathWorks, Natick, MA) scripts for this paper are available for downloading from http://pantherfile.uwm.edu/kravtsov/www/downloads/KWCT2015.

15 April 2015; accepted 6 November 2015
10.1126/science.aab3570