Has the intrinsic component of multidecadal climate variability been isolated?
A discussion of two recent papers: Steinman et al. (2015) and Kravtsov et al. (2015)

M.G. Wyatt
December 4, 2015 (*Note addendum and additional response from S. Kravtsov at end of essay.*)

Climate varies on time scales from years to millennia. Recent attention has focused on decadal and multidecadal variability, notably "pauses" in warming. The latest one - a slow down in warming since 1998 - has eluded explanation. One school-of-thought presumes external forcing (natural and anthropogenic) dominates the low-frequency signal; another casts internal variability as a strong contender. Numerous studies have addressed this conundrum, attempting to decompose climate into these dueling components (Trenberth and Shea (2006); Mann and Emanuel (2006); Kravtsov and Spannagle (2008); Mann et al. (2014); Kravtsov et al. (2014), Steinman et al. (2015); Kravtsov et al. (2015)). Despite many efforts, no universal answer has emerged.

Steinman et al. (2015) claim significant strides toward resolving the matter. They argue that they have identified an externally forced response, and per consequence, can estimate the observed low-frequency intrinsic signal. Using multiple climate models, they generate an average of all climate simulations. The resulting time series – a multi-model ensemble-mean – is their defined forced signal. Steinman et al. aver that this forced signal is completely distinct from intrinsic variability of all simulations within this suite of models. If this claim holds, then this model-estimated forced signal justifiably can be combined with observational data in semi-empirical analysis, which Steinman et al. use to expose the estimated intrinsic component within observed climate patterns.

But questions arise - about the modeled data and about procedure.

Output of climate-model simulations can provide insights into the observed climate response to external forcing. Steinman et al. use data from the collection of models of the fifth version of the Coupled Model Intercomparison Project (CMIP5). While a valuable resource; CMIP simulations have limitations. A major one is uncertainties surrounding model input: Specifying external forcing and parameterizations of unresolved physical processes is an inexact science; thus, in attempt to accommodate the speculated ranges and combinations of these factors, modelers script individual models within the CMIP multi-model ensemble with different subsets of such. Hence, each model generates a statistically distinct forced response. Deciding which forced response *might* be the "right" one is the challenge. Steinman et al. maintain that they have overcome this hurdle with a single forced signal that they claim represents the entire collection of CMIP models. In light of the individual model distinctions aforementioned, and the observation that each individual model generates its own unique forced response, the assertion of a single forced response, suitable for all CMIP models, is striking.

Much of Steinman et al.'s argument rests on residual time series. This is the information left-over from the time series of a climate simulation after the model-estimated forced response has been extracted from it. These time series can be used to ensure the forced signal is unbiased. The reason behind this assumption is that if the model-estimated forced response is completely disentangled from each model's intrinsic signal, and therefore unbiased, then the modeled intrinsic signals (the residuals) should be unbiased too; the residuals would be statistically independent of one another, i.e. not correlated. Steinman et al. claim to show the residuals are uncorrelated – a significant result, if true.

Steinman et al. generate the residuals by removing the forced signal from individual time series of *all* climate simulations, across all models and do so using two different methods – one involves differencing (subtraction) and the other, linear removal of a re-scaled forced signal via regional regression. Both operations yield similar results: Each leaves behind numerous residual time series, and each residual is taken to represent model-estimated intrinsic variability. Steinman et al. test correlation among the residuals: They invoke an indirect statistical test based on well-known properties of distribution of independent random numbers. The residuals are shown to be uncorrelated. This result suggests that the forced signal, indeed, is unbiased. Now the path is paved for Steinman et al. to use this signal to evaluate internal variability of observed regional climate patterns. Once they achieve this, they can use the identified intrinsic component to infer its potential relationship to the currently observed "pause" in surface warming. This all sounds promising. But is all as it seems?

The results seem counterintuitive, given fundamental differences among individual models. How could the forced signal truly be unbiased? This apparent puzzle motivated Kravtsov et al. (2015). They explore details of Steinman et al.'s methodology and design an alternate strategy to separate model-estimated signals.

In the big picture, Steinman et al. and Kravtsov et al. go about disentangling climate components similarly. Yet there are significant differences. Steinman et al. place emphasis on a *multi*-model ensemble. They consider simulated time series "in-bulk", so to speak, with no distinctions made among individual models. Kravtsov et al., instead, focus on the *individual* models of the multi-model ensemble; they look at simulations from one model at a time. This difference in approach produces different data sets of residuals. Specifically, Steinman et al. use a multi-model ensemble-average as their forced signal, and remove this forced signal from individual climate simulations *across all models* of the multi-model ensemble. The result is one data set of residuals for this multi-model ensemble. On the other hand, Kravtsov et al. generate two data sets of residuals *within each* of 18 *individual-model ensembles:* One data set [*1*] is derived from linear subtraction of a multi-model ensemble-mean from each simulation of a given model; repeated for 18 models. The second [*2*] is derived from subtraction of the single-model's ensemble-mean from each simulation of a given model; repeated for 18 models.

This latter approach [*2*] – subtraction of the single-model's ensemble-mean from a climate simulation - is a traditional method for decomposing simulated climate variability into its forced and intrinsic components, thereby generating a naturally unbiased signal. The reason for success in generating this naturally unbiased signal is traceable to a fundamental assumption: forcing subsets and physical parameterizations for a given single-model's ensemble of simulations are identical; any differences in the modeled realizations are due to differences in the model's intrinsic variability, most of these attributable to differences in initialization of each run, and therefore, differences that are uncorrelated. Thus, when the forced signal (single-model ensemble-mean) is subtracted (differenced) from each climate-simulation time series within a single-model ensemble, remaining residuals within this ensemble should be uncorrelated, or independent. And indeed, using a simple time series-correlation metric, Kravtsov et al. found this to be the case.

In contrast, when Kravtsov et al. linearly subtracted [*1*] the *multi*-model ensemble-mean (i.e. Steinman et al. forced signal) from the model simulations *within a single-model ensemble*, the resulting residual time series within the model's ensemble were significantly correlated [*3*]. These residual time series of a single-model ensemble are not independent. They share a common signature. They contain remnants of the multi-model ensemble-mean! Implications are significant: If the *multi*-model ensemble-mean – i.e. the Steinman et al. forced signal - produces a biased forced response for *single* CMIP5 models, how could this choice of forced signal credibly provide the *un*biased estimate of a forced response in observations?

All this brings us to a final curiosity: Steinman et al. demonstrated absence of bias in their forced signal. How did they do this, if indeed, the signal is biased? The answer lies in their procedure for ensuring statistical independence of model-estimated residual time series. Their procedure is flawed: due partly to their choice of forced signal – a multi-model ensemble-mean, and due partly to how the forced signal is removed from simulations. Recall, they remove the forced signal from individual simulations across *all* models, versus removing the forced signal from simulations *within* single-model ensembles.

Alluded to earlier in this essay was the procedure used by Steinman et al. to assess signal bias, or lack thereof. We revisit that procedure here. This simple method determines whether or not a number of individual residual time series share a common signature with one another. Bear in mind: if the forced response is completely separated from each simulation's intrinsic component, the numerous resulting time series of residuals will be uncorrelated with one another, i.e. statistically independent. Hence, when they are averaged together into a multi-model ensemble-mean of residuals, any differences among them would largely cancel out. The result of this is that the dispersion (difference from mean) of the multi-model ensemble-mean of residuals would be much smaller than the dispersion of the individual residuals. The procedure used to capture this relationship compares the actual dispersion (the variance of the ensemble-mean residuals) to the theoretical dispersion (the average of the individual variances of the residuals divided by the number of simulations). Due to the way variance is computed, the actual dispersion will be much smaller than the theoretical dispersion if the residuals are independent (uncorrelated). Steinman et al. applied this well-known method to their data set of residuals and found that their actual dispersion was, indeed, much smaller than their theoretical. It logically followed that their forced signal must be unbiased.

But it turns out there is a hidden glitch in this procedure that makes this result an illusion: Defining the forced signal in terms of a multi-model ensemble-mean and extracting it from individual simulations in-bulk, across all models, imposes an algebraic constraint on the residuals such that the ensemble-mean of residuals happens always to be zero, by mathematical construction. Due to this algebraic constraint, the actual dispersion will *always* be smaller than the theoretical one, and therefore the residuals always will appear to be uncorrelated, whether they are or not.

The trick in bypassing this constraint is to limit focus to residual time series within an ensemble of a single model. This is what Kravtsov et al. did. Had Steinman et al. also done this – i.e. removed their multi-model ensemble-mean from simulations exclusively within an ensemble of a single model, and tested the resulting residual time series for statistical independence using actual versus theoretical dispersion, as per Kravtsov et al., - they, too, would have found the residuals to be correlated (not independent).

Hence, through these slightly different methodologies, Steinman et al. and Kravtsov et al. come to different conclusions. Steinman et al. argue they have identified an unbiased forced signal, and with it, have identified the component of observed climate due to intrinsic variability. In contrast, Kravtsov et al. conclude Steinman et al. have not identified a forced response that is unbiased; that procedural artifacts only gave the illusion of such; and that successful disentanglement of climate components remains elusive.

A final point: Intrinsic variability indeed may damp the presumed anthropogenic signature of secular-scale warming in the currently observed "pause". Many have suggested such, Steinman et al. among them. Yet, without demonstration, conclusions on what may seem valid cannot be made. Thus, Steinman et al.'s argument that they have assessed the role of internal variability in the current "pause" is unsupportable. Their chosen forced signal - a multi-model ensemble-mean – fails to provide convincing evidence.

[*See addendum after appendix, acknowledgements, and list of references*.]
[*Also see S. Kravtsov's reply to Steinman et al.'s rebuttal (after addendum)*]

Appendix:

[*1*] Refers to the data set of residuals that Kravtsov et al. derived from linear subtraction of multi-model ensemble-mean from simulations *within an individual-model ensemble*. They used 18 individual-model ensembles (total of 116 twentieth-century simulations). Derivation of this data set adapts the Steinman et al. methodology in the following ways: one, a multi-model ensemble-mean of climate variability is identified as the forced signal; and two, this forced signal is separated out of each simulation, across *all* models. But in contrast with the Steinman et al. method, Kravtsov et al. do not evaluate independence of the residuals in-bulk; rather, they consider the resulting residuals in context of individual-model ensembles. Limiting analysis to single-model-residuals reveals that these residuals are, indeed, significantly correlated, and therefore not statistically independent. This method produces a biased signal.

[*2*] Refers to the data set Kravtsov et al. derived from differencing each individual-model ensemble-mean of climate variability (the forced signal) from each individual climate simulation within the simulation's associated individual-model ensemble. They used 18 single-model ensembles. Residuals within each individual-model ensemble are uncorrelated (independent). This method produces a naturally unbiased signal.

[*3*] These residuals are biased. Their "skew" is dominated by the difference between the naturally unbiased forced signal (individual-model ensemble-mean) and the Steinman et al. forced signal (multi-model ensemble-mean).

Acknowledgements: Feedback and editing suggestions from Sergey Kravtsov and Judith Curry streamlined the text of this essay; ensured its accuracy; and clarified its message. Their input is much appreciated.

References:

Kravtsov, Sergey and Spannagle, Christopher (2008); Multidecadal Climate Variability in Observed and Modeled surface Temperatures; Journal of Climate; volume 21; p1104-1121; DOI:10.1175/2007JCL11874.1.

Kravtsov, Sergey, Wyatt, Marcia G., Curry, Judith A., Tsonis, Anastasios A.; (2014); Two contrasting views of multidecadal climate variability in the twentieth century; Geophysical Research Letters, 41, 6881-6888; DOI: 10.1002/2014GL061416

Kravtsov, Sergey; Wyatt, Marcia G., Curry, Judith A., Tsonis, Anastasios A. (2015); Comment on "Atlantic and Pacific multidecadal oscillations and Northern Hemisphere temperatures"; Science, vol. 350, issue 6266; doi:10.1126/science.aab3570

Mann, Michael E. and Emanuel Kerry A. (2006); Atlantic Hurricane Trends Linked to Climate Change; EOS, Vol 87, No. 24, pages 233-244; Geophysical Research Letters,.

Mann, Michael E., Steinman, Byron A., Miller, Sonya K. (2014); On Forced Temperature Changes, Internal Variability and the AMO; Geophysical Research Letters ("Frontier" article); DOI:10.1002/2014GL059233

Steinman, Byron A., Mann, Michael E., and Miller, Sonya K (2015); Atlantic and Pacific multidecadal oscillations and Northern Hemisphere temperatures; Science, vol. 347, issue 6225; doi: 10.1126/science1257856.

Trenberth, Kevin E. and Shea, Dennis J. (2006); Atlantic hurricanes and natural variability in 2005; Geophysical Research Letters, vol. 33, L12704; doi:10.1029/2006GL026894.

***Addendum**: Efforts toward disentangling the two components that are assumed to generate low-frequency climate variability have been underway for about a decade (e.g. Trenberth and Shea 2006; Mann and Emanuel 2006). Work by Sergey Kravtsov, lead author of Kravtsov et al. (2015), is among them (Kravtsov and Spannagle 2008). The paper by Steinman et al. (2015) – discussed in this essay – is driven by the same goal. Yet, contrary to claims by Steinman et al., the stadium-wave studies (Wyatt, Kravtsov, and Tsonis 2012; Wyatt and Peters 2012; and Wyatt and Curry 2014) were not. The goal of the stadium-wave studies was <u>not</u> to disentangle external from intrinsic components; it was to analyze low-frequency behavior, regardless of its underlying temporal source.*

*The source of controversy is a step of methodology – linear detrending of raw data, a method commonly used to remove the secular-scale component in climate behavior, thereby highlighting shorter timescales of variability. Steinman et al. and Mann et al. have criticized this method as being ill-suited for disentangling the forced climate response from the component due to intrinsic variability. Authors of the stadium wave do not disagree. But, then again, disentangling the components of climate variability (component due to external forcing and component due to intrinsic variability) was **<u>not</u>** the goal of the stadium-wave studies. Highlighting multidecadal variability was, and linear detrending is a method well-suited for this goal.*

*The stadium-wave studies applied multivariate techniques to the detrended data. This was done in order to identify the shared time scale of variability among a collection of synchronized (matched rhythms, not matched phasing) indices. A multidecadally paced, hemispherically **propagating signal was thereby documented** (Wyatt, Kravtsov, and Tsonis 2012). Subsequently, and most intriguingly, CMIP3 modeled data were analyzed in an identical manner, resulting in <u>no</u> signal propagation (Wyatt and Peters 2012), underscoring **differences in outcome between modeled and observed data sets**, bringing into question data sets – a point further discussed in Kravtsov, Wyatt, Curry, and Tsonis (2014). And finally, a **mechanism** was proposed in (Wyatt and Curry 2014 (see also dissertation Wyatt 2012)), whereby the observed signal is shown to progress through a sequence of ocean, atmospheric, and Eurasian Arctic sea ice indices in a manner that gives a hypothesized physical basis to the hemispheric signal transfer – a feature captured in both observational and proxy data since at least the mid-1800s (Wyatt 2012), but not in modeled data.*

*It is a non-trivial point that the signal propagation, itself, is thought by the stadium-wave team to be intrinsic – rooted in boundary conditions, a key one being the unique juxtaposition of open ocean and sea ice in the Eurasian sector of the North Atlantic (see Wyatt and Curry 2014, section 4). <u>Yet</u>, it is significant to note that the **source for the tempo** at which the time scale of variability is paced (influenced by North Atlantic multidecadal variability (Atlantic Multidecadal Oscillation (AMO))), is thought likely to be a **combination of external forcing and intrinsic variability**. Thus, there is no disagreement between the stadium-wave studies and those by Steinman et al. on this speculation. Where disagreement lies is on the nature of response of a collective ocean-atmosphere-ice system to the forced and intrinsic components.*

In summary, the stadium-wave team makes no claim of having identified the *__source__* *governing the time scale of multidecadal climate variability or of the disentanglement of the two climate components generating the observed low-frequency signal. Steinman et al. have misinterpreted and/or misrepresented the stadium-wave studies. Refer to (Kravtsov et al. 2014); (Kravtsov et al. 2015); and related essays - all on this website.*

See: http://judithcurry.com/2015/12/15/20656/ for blog post on this topic.

Note that Steinman et al. replied to the technical comment (Kravtsov et al. 2015). Sergey Kravtsov's replied to this. His comments, posted below, can also be found at Judith Curry's website http://judithcurry.com/2015/12/15/20656/ and also on Sergey Kravtsov's website: https://pantherfile.uwm.deu/kravtsov/www/).

Sergey Kravtsov responds to Steinman et al. reply:

Steinman et al. claim that they avoided the issue of the algebraic constraint leading to the apparent cancellation of the "intrinsic" residuals in multi-model ensemble mean by using N-1 models to define the forced signal of the Nth model. However, doing so is really no different from using the multi-model mean based on the entire ensemble, since exclusion of one simulation will not affect the multi-model ensemble mean in any appreciable way. In fact, it is easy to show that the ensemble-mean residual time series in this particular case would be approximately proportional to the estimated forced signal – multi-model ensemble mean time series, – with the scaling factor involving 1/N, thus making its standard deviation much smaller than expected from 1/sqrt(N) scaling due to the cancellation of statistically independent residuals. Hence, the multi-model ensemble mean of the "intrinsic" residuals so obtained would have a negligible variance, but this will have nothing to do with the actual independence of the residuals (see Reference/Note 5 in Kravtsov et al.). Indeed, Kravtsov et al. demonstrated that these residuals are definitely well correlated within individual model ensembles, hence not independent. Steinman et al., in their reply, acknowledge the correlation, but still falsely claim independence.

Steinman et al. built the rest of their rebuttal on the fact that individual-model ensembles only have a few realizations, so the ensemble mean over these realizations, even if smoothed, will contain a portion of the actual intrinsic variability aliased into the estimated forced signal. Hence the estimated residual intrinsic variability will be weaker than in reality. This is a valid point, and it can be dealt with by choosing the cutoff period of the smoothing filter (5-yr in Kravtsov et al.) used to estimate the forced signal more objectively, for example by matching the level of the resulting residual intrinsic variability in the historical runs with that in the control runs of CMIP5 models. However, this issue is only tangentially related to the implications and fundamental limitations of using the multi-model ensemble mean to estimate the forced signal outlined in Kravtsov et al. In particular, if one would use multi-model ensemble mean to define an intrinsic variability in any one model, this variability would have a much larger variance than the intrinsic variability of this model in the control run; this larger variance would be dominated by the individual forced signal bias for this model.