Geosci. Model Dev. Discuss.,
doi:10.5194/gmd-2016-285-AC2, 2017


Geoscientific Model Development Discussions
Open Access
EGU



# *Interactive comment on* "Skill and independence weighting for multi-model assessments" *by* Benjamin Sanderson et al.

**Benjamin Sanderson et al.**

bsander@ucar.edu

Received and published: 4 April 2017

Thanks to the reviewer for his thoughtful reading and suggestions. We lay out below our thoughts in regard to his review, and how our paper relates to the author's work on the topic. We attach a revised version to address the reviewer's concerns and to better represent his own work on the topic.

*Fig. 4a shows the improvement over the sample mean as a function of their key tuning parameter for this historical data. This figure indicates that the optimal parameter value for the combined metric is between 0.3 and 0.5 (even though the text just gives it at 0.5). (The authors need to explain how they got 0.5 from Fig 4a rather than 0.4 or 0.3).*

The historical RMSE score isn't the only consideration, i.e. we don't only use Fig. 4a in our selection of the parameter - the value chosen is 0.8 or 80 per cent of the best-



Printer-friendly version

Discussion paper

model/obs distance. We did state that the lowest in-sample score was achieved with a value of approximately 0.5, but the next paragraph notes that this isn't how we choose our metric because choosing based on in-sample data only would lead to an overly confident constraint. Sorry for this confusion, we've reworded the first paragraph to make this clearer. We agree that the curve minimum in 4a is closer to 0.4 and have updated the text to reflect this, but note this was just an observation, this value is not used in any further analysis.

The two other factors considered are the out of sample (2080-2100) skill in Fig. 4b and the risk that our weighting would produce a distribution which increased the risk of the true model falling outside the weighted distribution. Hence - if historical RMSE was the only concern, we would choose a value of 0.4 - which would give us a better in-sample RMSE. The value of 0.8 is chosen such that the risk of overfitting is minimized, while still allowing for some moderate increase in weighted in-sample RMSE score.

*Fig. 4b shows how the tuning parameter actually affects the 2080-2100 forecast accuracy, not for the combined metric but just two of the variables within the metric. Comparing Figure 4b with 4a shows that if one had used a good value of the parameter for the combined metric from the historical data, 0.4, say, the weighted multi-model mean would actually give a similar or less accurate precipitation and temperature forecast than the simple sample mean. This inability of the weighting method to produce significant forecast improvements when tuned against historical observations suggests C2 that the proposed method may be of little value.*

As noted above - overfitting the historical RMSE would reduce the out of sample skill, and we specifically don't do that for that reason. Hence, a less aggressive weighting was used - informed by Figs. 4b and 4c. Using the final value of 0.8, there is a small increase in out of sample skill - but we agree, it's a minor increase. But, we also don't find this particularly surprising - if there existed strong relationships between the mean state and the future temperature or precipitation changes, these would be exploitable emergent constraints in their own right. The literature has demonstrated





Printer-friendly version

Discussion paper

consistently that these constraints are rarely found in the CMIP archive. The fact that CMIP5 models on average agree better with observations than CMIP3 has not resulted in a more narrow projection range.

Our defense of the technique is that it provides a simple way to downweight clear model duplication, and relatively poor models in the archive. This may or may not result in a more accurate ensemble predictions, but there is no way to know whether a biased ensemble provides a biased projection that to see whether the weighting makes a difference. As we note, the actual CMIP archive has a tendency to have more replicates of models which exhibit lower RMSEs, there aren't many examples of models which exhibit huge biases both in the present, and there are no clear emergent constraints on future change - so the effect of the technique on CMIP5 is subtle because the model average happens to be almost optimal.

Our argument is that our method allows an analysis future-proofed against future ensembles with very poor models or with large numbers of replicates. If a group submits 1000 versions of the same model to CMIP6, our method would do a defensible job of allocating an appropriate amount of weight without modification. Similarly, if someone submitted a perturbed physics ensemble containing some model versions which were completely unlike Earth in the present, the presented method would downweight them appropriately. We agree that the use of the AB15 method has relevant merits for model weighting, but our needs in this case were specific, following a request for one set of positive model weights which could be used for further analysis by the authors of the report to address model skill and interdependence in a simple way.

*Nevertheless, there is merit in other aspects of the paper and with major revision; the paper could make a useful contribution to the field.*

***Specific Comments*** *The poor climate projection results obtained from the authors' proposed method when tuning using pseudo historical observations are in contrast to the findings of work I have been involved in. Specifically, in similar tests to those of*

*Sanderson et al, Abramowitz and Bishop (2015, J. Clim) (AB) obtained average reductions in the root mean square distance from the out-of-sample truth greater than 30 percent when using the climate ensemble member weighting method of Bishop and Abramowitz (2013, Climate Dynamics) (BA). The current version of the paper lacks any reference to AB. Furthermore, on lines 67-68, it dismisses BA's approach as being undesirable for their North American application. This is incorrect. Small root mean square forecast errors is universally accepted as a desirable aspect of a forecasting scheme. AB showed that relative to the root mean square error of the uniformly weighted ensemble mean, the reduction in root mean square forecast errors due to the BA weighting method is profound. Furthermore, their method can easily be "geographically focused" for regions such as North America. As such, I strongly encourage the authors to revise their draft so that it acknowledges BA's approach as potentially useful for North America and discusses the positive results of AB. Obviously, AB considered differing metrics to Sanderson et al. so no apples-to-apples comparison can be made between AB's results and the results of this paper but AB's work needs to be recognized and not dismissed as undesirable because of the BA method's use of metamodels. Each of BA's metamodels is a linear combination of the original models constructed so that the weighted mean formally minimizes error variance; and the BA ensemble variance is equal to this minimal value of the error variance. One needs to recognize that each raw climate model is itself a "meta Earth system" that is a crude approximation to the real Earth system.*

We now devote a number of paragraphs to the description of the reviewer's 2013 and 2015 papers. AB15 is an interesting and novel framework for ensemble analysis, but it could never have been an option for this particular application because the request for the National Climate Assessment was specifically for one set of model weights which reflected model skill and independence. The weights were then passed to the author team, who conducted individual analyses for the NCA. As such, we were structurally constrained to produce a product which could be simply used by the author teams. A single set of weights could be incorporated fairly simply into the large number of pre-

defined analyses which go into such a report (which is for general public consumption), whereas a transformation into statistical meta-models which do not, in themselves, follow physical laws would have been practically impossible to implement by the author team.

But - we do note that the comparison of 30 per cent reduction in out of sample truth is not comparing like with like. Firstly, the 30 per cent out of sample skill increase referred to in AB15 is the absolute difference between the mean state of the 'perfect' model and the optimized ensemble regression prediction in a period out of the training period. The out of sample skill in 4b in this paper is the skill in predicting the anomaly between present day T/P and the future. Part of the skill in AB15 comes from persistence of mean state bias - which is taken out of our test.

Secondly, although AB15 goes to some efforts to remove duplicates in their perfect model tests - they are not extensive. For example, AB15's "independent" test ensemble contains both CESM1 and NorESM1, and HadGEM2 and ACCESS - which both contain near replications of the atmospheric models. In this study, we have gone to significant efforts to remove any duplicates from our perfect model test, which would have trivially increased our out of sample skill.

*It is true that AB15's metamodels are unrealistic in that, for example, they do not obey conservation laws for energy and mass. However, they are more realistic than the original models in the sense that their statistical relationship to historical observations is more like that of an ensemble of perfect models (replicate Earths) than the original models.*

We would argue AB15's historical RMSE score is smaller by construction (there is no linear combination of models which could have a smaller RMSE), and the future reduction in anomaly projection error is not shown in AB15. But given that it is not empirically clear that one model subtracted from another is a physically meaningful quantity, only future anomaly error reduction in a true perfect model test where no close relatives of

the perfect model exist in the archive would constitute definitive evidence of greater skill. It could be argued that the any average of several models is also not necessarily physically meaningful because any combination of models no longer follows conservation laws, but a weighted average of models has a simple interpretation: a combined measurement of a number of models, weighted by their trustworthiness. Formulating the problem as a regression equation allowing negative coefficients though creates a more difficult product to interpret.

Given more models than degrees of freedom in the CMIP5 dataset, one could produce a near-perfect reproduction of the observations. Hence in order to be sure that AB15 is not subject to overfitting, it would be necessary to demonstrate that the degrees of freedom in CMIP models significantly exceed the number of fitted points. For a simple spatial field like temperature - where a few spatial modes can well define the response patterns of different models in the archive, this may not necessarily be the case.

*Having found rather poor forecasting results when using weights derived from pseudohistorical data, Sanderson et al. then consider weights that are tuned for model forecast data so that, on average, they deliver a weighted mean that is as close as possible to the 2080-2100 state of a climate model excluded from the set of ensemble members used for the forecast (Fig 4b). In statistics, such "in-sample" statistical tests are viewed with suspicion because of the possibility of overfitting.*

In our study (in contrast to AB15), we have only one parameter - so we don't have the ability to overfit in the regression sense of the word. We are not fitting to the future data directly, we are just reducing the degree to which the present day values can constrain the data if in the perfect model weighted average prediction of future anomalies can be demonstrated to be overconfident. Fig 4b is thus a diagnostic to show that if we had chosen an optimal value of the skill radius to maximise in-sample skill, then this would be non-optimal for out of sample skill. But the metric itself used to determine the parameters only considers historical data.





Printer-friendly version

Discussion paper

*An additional concern about this approach is that it would be impossible to apply it to real observations (unless one waited until 2100 when the data would be available). One is left having to justify the approach on the assumption that the climate models are producing realistic future climate data. In contrast, if as in AB and BA, one demonstrated improved forecasts using historical observations, there would be much less room for argument about the realism of the data available for tuning.*

We do apply the approach to observations - the constraints are entirely based on historical observations. We only use the future data in the models to assess how strong the costraints on past performance should be in general. A regression-based approach such as AB2015 has the capacity for overfitting, if the number of degrees of freedom exceed the number of models. Our technique calibrates a single parameter - which represents the degree to which historical data should weight a given model's future projection. The 2100 skill is a diagnostic, not a component of the weight and the method cannot 'fit' the combined model result to the 2100 data. Figure 4b simply says "if we over-constrain the models to their present day performance, then our prediction of future anomalies becomes less accurate". Therefore, we don't need the 2100 data from the real world to be able to use our method - we only use historical data - but 4b tells us that we should weaken that constraint from what we would have inferred from past performance alone. So we would argue that 4b is the opposite of overfitting, it explicitly weakens our constraint to to ensure against overfitting.

*The revised paper needs to clearly address these concerns. In addition to the aforementioned issue, the point by point comments below highlight other major and minor issues that, if addressed, would improve the paper.*

***Point by point and technical comments***

*1.Line 67-68. See above comments.*

We have significantly expanded this discussion in the light of the reviewer's comments.

*2. Sentence from line 74-76. Suppose that one had two simulations from a perfect model and that each was started with a different initial condition. In this case, the model for each of the simulations is the same even though, because of the chaotic nature of the Earth-system, the state estimates obtained will have differences. It can be shown that the mean of these two random perfectly realistic states would have considerably less distance from another perfectly realistic state (Bishop and Abramowitz, 2015). Hence, not including the second ensemble member simply because the model that produced it was identical to the model used for the first model would reduce the utility of the ensemble. Thus, this idea and its incorporation into the weighting scheme does not seem to be well justified. Perhaps the authors assumed that over a long enough averaging period the time-means of the two simulations would be identical. Long range modelling studies of low-frequency variability such as that of James and James (1989, Nature 342, 53 – 55) do not support this assumption. The revised paper should comment on this issue.*

This point is well taken, but it does not address the key aspect of the CMIP5 ensemble which we are trying to address - that all of the models are not perfect, and that some of them are near replicates of each other. Our technique does not throw out any models - but it allocates approximately equal fractional weights to near-identical models.

The relevant thought experiment is the following. Let's assume we have 3 models, 2 of these are structurally identical to each other, and the third has a different structure. Both of the structurally identical models have some underlying bias in their climate attractor, and the third model has a different bias - but the bulk errors are comparable.

In this case, knowing the above information - we would argue that the correct distribution of weight is $\frac{1}{4}$ for each of the structurally identical models and $\frac{1}{2}$ for the unique model, and this is the solution solved for in this paper. This conclusion has nothing to do with averaging periods (although clearly, the shorter the time series, the noisier the result will be).

Our previous work (Sanderson et al (2015b)) shows that the inter-model distances due to internal variability are an order of magnitude smaller than the differences between structurally dissimilar models in the CMIP archive, when evaluated using a similar metric to that used in this paper using 30 year climatological means. As such, the effect of bias due to model replication is well resolved in the context of noise generated by internal variability.

*3. Section 3. Please add more details about the length and temporal filtering of the data set used to create the distance matrix.*

We added the following paragraph: " Data from each model is taken from the first available initial condition member of each model's historical contribution to CMIP5. Data from years 1976-2005 are used from each model, averaging all years to form a monthly climatology. Data from the observations are monthly climatologies averaged from all available years within the 1976-2005 window."

*4. Line 91-92 and Table 1. Extreme values such as 'coldest day' are highly prone to large variations that are simply due to random sampling rather than any error in the distribution being sampled. One can easily prove this to oneself by sampling a normal distribution of 20x365 random normal numbers and seeing how much the minimum value changes. I did 12 such trials and found values ranging from -3.29 to -4.25. In contrast, if I look at the variation of standard deviations for 12 such trials I get values with the very small range of 0.98 to 1.01 – only 2 percent variation. By rewarding with high weights ensemble members that happen, by pure chance, to get extrema correct, you may be compromising the potential performance of your ensemble weighting technique. Why not use a standard deviation metric instead?*

Using a standard deviation assumes a normal distribution which is inappropriate for assessing the properties of the tail of the distribution. It also assumes that the distribution is bounded - and climate variables are not. The CSSR/NCA requires an assessment of extreme model behavior, and we use metrics from a well-established community

to form the statistics (we use the methodology laid out in Sillmann et al (2013), which shows such statistics are well sampled for a 20 year climatology - and we use 30). Note also that data at high temporal resolution is not always publicly available, whereas the standardized extreme indices are readily available for models and observations.

*5. Caption of Fig 3. What does NCA4 stand for?*

Expanded to the full name of the report.

*6. Subsection 3.5. It seemed that you held the independence weights constant for section 3.5. Please be clearer about how these were combined with the skill weights for the experiments reported on in Subsection 3.5.*

Text added to the paragraph: "In Figure 4(a), we use the uniqueness parameter Du determined in section 3.4 and sample a range of Dq."

*7. Legend of Figure 4a. Are the "ta" and "tas" mentioned in this legend respectively the same as the "T" and "TS" mentioned in Table 1?* The revised paper needs to ensure that Table 1 is consistent with this legend and vice-versa. Table 1 is now consistent in abbreviations.

*Also, on my copy of the paper, C5 in Fig 4a it was extremely difficult to tell which line corresponded to which variable. It would be clearer if, in addition to color, you used shapes (triangles, boxes, diamonds, asterisks, etc) to help distinguish which line belongs to which variable.* The figure has been reformatted for clarity as the reviewer suggests.

*8. Line 165. Here you state that Figure 4a suggests to you that 50 percent (0.5) minimizes forecast error. To my eye it looks like 0.4 or 0.3 minimizes forecast error. Please give more details about how you came up with the 50 percent value.*

We agree - we've changed the text. As explained above - this value was just an observation from the graph, it was not used in any part of the further analysis.

*9. Line 191. Change "averages" to "averaged"*

Done

*10. Line 198. Please provide more information about how you "skill weighted the ensemble". Does this create a new ensemble? How do you assess whether the truth lies within or outside of this skill weighted ensemble? I am unable to comment on any aspect pertaining to Fig 4c because of my uncertainty about what you actually did.*

We have considerably increased the length of this discussion.

*11. Weight normalization. The text is somewhat unclear about where and when the weights are normalized so that they sum to 1. Please be clearer about this. An equation stating exactly what you did would be helpful.*

Added equation 7.

*12. Figure 5. I like the idea of excluding similar models for the "model as truth" experiments. This option was not investigated by AB. Do your results change much if you don't exclude any models?*

Quite a lot, depending on the model and variable - not excluding clear replicates like NorESM/CESM tends to produce out-of-sample anomaly projection skill which is artificially high in the model as truth experiments. Keeping all members for the perfect model case therefore reduces the apparent out-of-sample skill a lot. Figure 1 (in this response) shows the equivalent of Figure 4b in the main document without prefiltering for near neighbours. The method would suggest a "model-as-truth" best average score of about 30 percent below the simple multi-model mean for precip, and 15 percent for temperature. I.e. It would give too much confidence in the out of sample skill.

*13. Line 216 – 218. State quantitatively what values are used. The previous sections used a whole range of values so it is unclear what precise values were finally chosen.*

Done.

*14. Line 430: Change “not trivial matter” to “not a trivial matter”*

Done

**References**

Abramowitz, G., and C. H. Bishop. "Climate model dependence and the ensemble dependence transformation of CMIP projections." Journal of Climate 28.6 (2015): 2332-2348.

Bishop, Craig H., and Gab Abramowitz. "Climate model dependence and the replicate Earth paradigm." Climate dynamics 41.3-4 (2013): 885-900.

Sanderson, Benjamin M., Reto Knutti, and Peter Caldwell. "A representative democracy to reduce interdependency in a multimodel ensemble." Journal of Climate 28.13 (2015): 5171-5194.

Sillmann, J., et al. "Climate extremes indices in the CMIP5 multimodel ensemble: Part 2. Future climate projections." Journal of Geophysical Research: Atmospheres 118.6 (2013): 2473-2493.

Please also note the supplement to this comment:
http://www.geosci-model-dev-discuss.net/gmd-2016-285/gmd-2016-285-AC2-supplement.pdf

---

Interactive comment on Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-285, 2016.



Printer-friendly version

Discussion paper

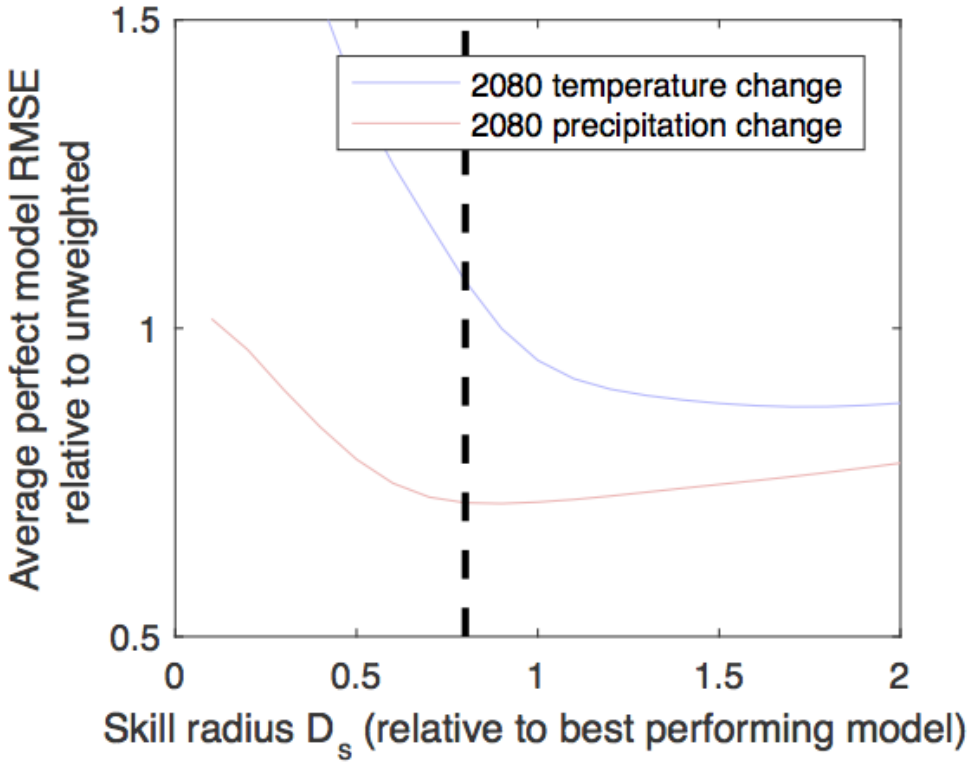


**Fig. 1.**



Printer-friendly version

Discussion paper