Geosci. Model Dev. Discuss.,
doi:10.5194/gmd-2016-285-RC2, 2017
© Author(s) 2017. CC-BY 3.0 License.

Geoscientific Model Development Discussions
Open Access
EGU

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

C. H. Bishop (Referee)

craig.bishop@nrlmry.navy.mil

Received and published: 6 March 2017

General Comments

Climate models mathematically express our understanding of the factors governing weather and climate. The CMIP5 multi-model archive pools together climate projections from leading weather and climate forecasting centers. Taken together, they represent an unprecedented concentration of mankind's knowledge and climate prediction capabilities. However, not all of the models are of the same quality and some of the models are very similar to each other. How then could the information from the individual models be combined to minimize the uncertainty in climate change uncertainty? One option is simply to take the unweighted sample mean of the ensemble of predictions. An alternative option pursued by Sanderson and co-authors in this paper is

Printer-friendly version

Discussion paper

CC BY

that of assigning weights to the ensemble members based on a combination of their skill and uniqueness characteristics. These weights are then used to create weighted ensemble means and uncertainty estimates for the weighted ensemble mean. Given the key role of climate change prediction to effective climate change mitigation, adaptation and reduction activities, studies such as these aimed at reducing climate change uncertainty are vitally important.

A major and compelling reason for widespread confidence in any prediction or theory is the accuracy with which it has been able to predict events in the past. As such, historical observations are a natural data set to try and tune the weights associated with any ensemble weighting scheme.

To indicate the possibility that the authors proposed weighting approach would lead to more accurate climate change projections if tuned against historical observations, the authors remove one of the climate model simulations from their ensemble and treat it as if it were the unknown but historically observed truth. With this approach they can generate pseudo-historical observations and then optimize their weighting approach for that historical pseudo-data. They then checked whether the weights derived in this manner led to more accurate forecasts of the truth-model's state from 2080-2100. 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). 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

Printer-friendly version

Discussion paper

CC BY

that the proposed method may be of little value. 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% 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. It is true that Bishop

Printer-friendly version

Discussion paper

CC BY

and Abramowitz'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.

Having found rather poor forecasting results when using weights derived from pseudo-historical 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. 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. 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.

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 con-

CC BY

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

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% 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?

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

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.

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. Also, on my copy of the paper,

CC BY

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.

8. Line 165. Here you state that Figure 4a suggests to you that 50% (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% value.

9. Line 191. Change “averages” to “averaged”

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.

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.

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?

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.

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

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

CC BY