Geosci. Model Dev. Discuss., 8, C926–C932, 2015
www.geosci-model-dev-discuss.net/8/C926/2015/

Geoscientific Model Development Discussions

Interactive comment on "A new ensemble-based consistency test for the Community Earth System Model" *by* A. H. Baker et al.

Anonymous Referee #1

Received and published: 1 June 2015

1 Overview:

Review of "A new ensemble-based consistency test for the Community Earth System Model" by Baker et al.

Baker et al. present a new method of determining if a new climate run is statistically distinguishable from an ensemble of "trusted" simulations. The method is simple and should be easily extendable to other climate models. The writing is generally clear and precise. I think the manuscript could be published provided the authors satisfactorily address a few major questions about the methodology.

1. Does the paper address relevant scientific modelling questions within the scope

of GMD? **Yes**

2. Does the paper present a model, advances in modelling science, or a modelling protocol that is suitable for addressing relevant scientific questions within the scope of EGU? **Yes**
3. Does the paper present novel concepts, ideas, tools, or data? **Yes**
4. Does the paper represent a sufficiently substantial advance in modelling science?
5. Are the methods and assumptions valid and clearly outlined? **Mostly**
6. Are the results sufficient to support the interpretations and conclusions? **Mostly**
7. Is the description sufficiently complete and precise to allow their reproduction by fellow scientists (traceability of results)? In the case of model description papers, it should in theory be possible for an independent scientist to construct a model that, while not necessarily numerically identical, will produce scientifically equivalent results. Model development papers should be similarly reproducible. For MIP and benchmarking papers, it should be possible for the protocol to be precisely reproduced for an independent model. Descriptions of numerical advances should be precisely reproducible. **Yes**
8. Do the authors give proper credit to related work and clearly indicate their own new/original contribution? **Yes**
9. Does the title clearly reflect the contents of the paper? The model name and number should be included in papers that deal with only one model. **Yes**
10. Does the abstract provide a concise and complete summary? **Yes**
11. Is the overall presentation well structured and clear? **Yes**

12. Is the language fluent and precise? **Yes**

13. Are mathematical formulae, symbols, abbreviations, and units correctly defined and used? **Yes**

14. Should any parts of the paper (text, formulae, figures, tables) be clarified, reduced, combined, or eliminated? **No**

15. Are the number and quality of references appropriate? **Yes**

16. Is the amount and quality of supplementary material appropriate? For model description papers, authors are strongly encouraged to submit supplementary material containing the model code and a user manual. For development, technical, and benchmarking papers, the submission of code to perform calculations described in the text is strongly encouraged. **Yes**

2 Major comments:

Assumption of non-skewed distributions
One of the metrics used for determining if a simulation is different from the ensemble is the Z-score. The Z-score assumes that the distribution is non-skewed. It's not clear to me that the distributions of all the variables would be symmetric. However, I can't say for sure if the distributions would be symmetric because the authors do not list the variables they are testing (see minor comment #1). I suspect that some of distributions would be skewed because many variables in atmospheric science tend to be bounded. For example, precipitation cannot be negative and would be more likely to follow a lognormal (or another skewed) distribution. The authors should justify this assumption or apply the proper data transformation.

T-scores vs. Z-scores
It seems that the use of T-scores and t-tests would make this work more generalizable. The use of Z-scores assumes that you have sufficient samples while the T-scores do not make this assumption. I don't think this would impact the results because the authors have employed a large ensemble (so the t-distribution should be approximately equal to a normal distribution), but it would make the equations more robust for others to use.

Is PCA necessary?
It's not clear to me why PCA is necessary. The authors claim to use it because "determining whether or not the climate in the new run is consistent with the other ensemble data based on the number of variables that fall within the distribution. . . is difficult without a linearly independent set of variables." The original 134 variables may not be linearly independent but they will span the whole space. Using 50 PCs won't span the whole space. Why not just use the original 134 variables? Reducing to 50 PCs doesn't seem like it would provide much of a computational benefit – because the system is already quite small (151×134). How does just looking at the 50 leading PCs help the analysis?

Why not, instead, use all 120 PCs in the analysis and report the percent of variance that is explained by the PCs that fail the test (e.g., Mira had X PCs fail but those PCs only explain XX% of the variance). The authors could then define a pass/fail cutoff based on the percentage of the climate variability that is explained – which seems less arbitrary than defining a cutoff of 3 PCs. This would also allow the authors to ensure the leading PCs are captured, which are presumably more important because they're the dominant climate "features".

3 Minor comments:

What variables?
There are many occasions in the manuscript where the authors say, for example, "there are 151 variables in CAM and the coefficient of variation for each variable is well under five percent, save for two variables that are known to have large distributions across the ensemble." Which two variables have large variation? Please state what variables you're using so the reader can follow along. This is also important for the major comment on "skewed distributions".

Averaging time
How long do you average the results over? I see the authors claim that the parameterizations in CAM make it ill-conditioned, but surely they must throw out a few of the initial days. I can't imagine that an inconsistent (climate-changing) simulation would be different from a consistent simulation after a few time steps for a perturbation of $\mathcal{O}(10^{-14})$. Is there any "spin-up" time?

Perturbations
Page 3831, Lines 16-17: Are you doing global or local perturbations (ie. are you perturbing the entire temperature field or just one grid cell?).

Incorrect citation formats
There are numerous instances of improper formatting for the citations. For example, on Page 3828, Lines 22-24 the authors write, "Some of the difficulties caused by differences due to truncation and rounding in climate codes that result in non-BFB simulation data are discussed in (Clune and Rood, 2011)." The correct citation format is the non-parenthetical form: "Clune and Rood (2011)". I have listed a few instances I found in the paper:
- Page 3827, Line 18: (Easterbrook and Johns, 2009)
- Page 3828, Lines 23-24: (Clune and Rood, 2011)
- Page 3828, Lines 26-27: (Rosinski and Williamson, 1997)

- Page 3831, Lines 2-3: (Rosinski and Williamson, 1997)
- Page 3831, Line 13: (Kay et al, 2015)

Better roadmap for the methods
As I was going through the methods section, I thought the Z-scores were the primary method for determining consistency. It wasn't until the results section that I realized the authors were primarily relying on the PCs for evaluating consistency. For clarity, it would help if the authors mention they have two methods for evaluating the consistency in the first paragraph of Section 3.

Validation vs. Evaluation
It seems that "evaluation" would be a better term than "validation" to use in the introduction section because one cannot demonstrate predictive reliability with a climate model (e.g., Orsekes et al., 1994; Orsekes, 1998). A climate model, inherently, produces possible realizations of the climate system. It would be foolhardy to attempt to "validate" a climate model. However, one can "evaluate" a climate model and determine what features of the model are reliable and useful for answering scientific questions and decision-making.

Page 3829, Lines 5-6: "The result is that the tolerance for rounding accumulation growth are exceeded within the first few time steps." Really? That seems (surprisingly) fast.

Page 3831, Line 20: Should say "computationally", not "computational".

Page 3840, Line 20: What is "ne= 30"?.

4 References:

Orsekes, N., K. Shrader-Frechette, K. Belitz: Verification, validation, and confirmation of numerical models in the earth sciences. *Science* **263**, 641-646, 1994.

Orsekes, N.: Evaluation (not validation) of quantitative models. *Environ. Health Perspect.* **106**, 1453-1460, 1998.

Interactive comment on Geosci. Model Dev. Discuss., 8, 3823, 2015.