Atmos. Chem. Phys. Discuss., 9, C2378–C2382, 2009
www.atmos-chem-phys-discuss.net/9/C2378/2009/

Atmospheric Chemistry and Physics Discussions

Interactive comment on "Comment on "Quantitative performance metrics for stratospheric-resolving chemistry-climate models" by Waugh and Eyring" *by* V. Grewe and R. Sausen

Anonymous Referee #1

Received and published: 30 June 2009

This paper by Grewe and Sausen (GS09) is a comment on an earlier ACP paper by Waugh and Eyring (WE08) that introduces a method to define a quantitative performance metric ("grade") of chemistry climate models (CCMs). While the authors generally acknowledge the effort to define such "grades" they point out that the method proposed in WE08 is not statistically sound.

Part of their criticism were already expressed in the open discussion phase of WE08, however, the paper of WE08 was not revised in a way that these criticisms were accounted for.

The paper of WE08 was the first to my knowledge that aimed at quantitatively compare and grade CCM models. This was seen as a major step forward in assessing the quality of CCMs. Briefly, in WE08 first a model performance of a relevant quantity was determined by comparing with observations, then a model grade is constructed from many relevant quantities, and thirdly, this model grade is used to construct a weighted consensus for model predictions of many models. In the current paper, Grewe and Sausen convincingly point out the shortcomings of the method, especially the first step of assigning a grade to a model with regard to a certain quantity. This paper is well written and should be published in ACP with minor revisions that are listed below.

General points and remarks

1. From reading this paper (and WE08), it becomes clear, that the variability (standard deviation, measurement error) of an observable is very important for the grading procedure. Depending on which diagnostic a grading is based, it may be very difficult to adequately address the statistical error and uncertainty values to it. This is required for getting a robust evaluation of the diagnostic. More precisely, the dominant source(s) contributing to the standard deviation of the observable must at least be known.
   E.g. the standard deviation in the HALOE climatology used in WE08 is a composite of inter-annual and geographical (zonal) variability at the points covered by HALOE. It does not include the experimental (retrieval) error. However, it is implicitly assumed, that the given variability is the main contribution to the standard deviation. This assumption needs to be verified, which may not be easily possible, as especially for satellites the retrieval error etc. are not exactly known. In other words, the exact definition of the variabilities discussed in this paper and the resulting grade definition may not be possible without some educated guesses. This point may be mentioned in the paper, it does however not reduce the relevance of the given criticism brought up by GS09.

ACPD
9, C2378–C2382, 2009

Interactive Comment

2. On p. 14145, l. 9ff, GS09 point out that the sigma of WE08 is either based on inter-annual variability or measurement uncertainty:
Different aspects may contribute to this variability. E.g. for the first diagnostic of WE08, mean south pole temperature (SON, 60-90N, 30-50hPa), there is a variability within the given time, latitude and pressure range, a year-to-year variability and the "measurement uncertainty" (that again is a model uncertainty of the ECMWF model). If a model is somehow forced ("nudged") by observations of QBO phase, winds or temperatures, the model may reproduce the reality, potentially not because the model is good but because of a good forcing. Therefore the rationale of model grades seems to be best suited for free runníng models.

3. The question, what observational error should be tolerated (p. 14145, l. 28ff) is very important. If an observation has a very large error, i.e. little valuable information can be drawn from this observation the grade after WE08 would be large. On the other hand, if there is a very accurate observation, already a rather small offset between the model and the observation may result in a low grade. One may decide in certain cases independently from the statistics, which deviation from the observations should be tolerated.

Minor points

1. p14142, l. 13, abstract, since...drastically
This is not a good justification for the need to include confidence intervals into the grading. I would rather expect a justification like "since otherwise a perfect model may get a low grade" or so...

2. p14144, l. 1: This is not a correct implication. In principle the model could represent a test object correctly by accident, without having built in the correct key processes.

ACPD
9, C2378–C2382, 2009

3. p14144, l. 17-24: These mentioned errors with respect to the representativeness of some (satellite) data can in principle be accounted for when comparing simulations with observations. It may not often being done in publications, however there are studies in which observations are compared with simulations sampled at exact location and time and the model output is degraded to the experimental vertical resolution.

4. p. 14145, l. 19ff and p. 14146, l. 5-7, Gaussian distribution
One can not always expect Gaussian distributions for observables. Later on in the paper, it is shown that the arguments may be very similar for a uniform distribution. However, a hint could be added at this stage, that Gaussian distribution is very typical and other distributions will also be discussed.

5. p. 14146, l. 8, figure 1: It is somewhat confusing that the variable is called x and also the "models" are called X and Y. Please use another letter for the variable.

6. p. 14149, l. 9f, This argument is not clear to me. The grade of a model used as an observable. From fig 1 it is clear, that the expectation E(G) cannot be 1 since the two representations of the model get grades below 1. So it is clear that E(G) cannot be 1, but the reason of the formula is not clear.

7. p. 1151, fig. 4, The 5% and 1% percentiles are shown to determine the statistical error for the given grade. It means that two identical models have in 95/99% of the cases a grade difference less than this number. Evenly interesting would be the "typical grade difference" which is the median or the 50% percentile.

8. p. 14151, l. 14, Do you mean: The maximum value of α_{mod} and β_{mod} is 2?

9. p. 14151, l. 19, Have all iteration been calculated with N=M=10?

10. p. 14153, l. 13: The suggested grading called "s-grading" accounts for the statistical error of an observable. It is not clear to me how exactly this quantity is

ACPD
9, C2378–C2382, 2009

defined in the paper. One can think of different definitions. Please clarify.

11. The paper is written in a consequent mathematical way which I do appreciate. However, in some cases a few words could be added or changes to the figures could be made to make the statements more illustrative, e.g.
draw 3 dashed vertical lines in figure 2 for the mean grade, 5% and 1% percentile. This would optically support the statements made in the text and give a better link to figure 3a-3c.

Typographical corrections

1. p. 14143, l. 12/l. 14 add comma: In Sect. 3, we define
2. p. 14155, l. 17 suggest
3. p. 14155, l. 18 differs

Interactive comment on Atmos. Chem. Phys. Discuss., 9, 14141, 2009.