Atmos. Chem. Phys. Discuss., 9, C3463–C3467, 2009
www.atmos-chem-phys-discuss.net/9/C3463/2009/

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 #2

Received and published: 2 August 2009

Recommendation: Not publishable in its present form.

This paper provides a critique of the model grading exercise of Waugh and Eyring (2008), hereafter WE08. It does so by Monte Carlo statistical tests using simulated perfect and imperfect models and observations. The paper does not take issue with any particular result of WE08, but argues that the statistical uncertainty already recognized by WE08 was not propagated into the model grading (in the sense that no confidence intervals for the grading are given), and that allowing for realistic levels of model or observational error would widen these confidence intervals to the point that it would be

very hard to distinguish the skill of any model from another, or refute the hypothesis it was perfect, at 95% or especially at 99% confidence.

I appreciate what the authors are trying to do here. It is important that any quantitative performance metric be statistically reliable, and as with any measurement – for a grade is a 'measurement' – confidence intervals simply have to be provided. Nobody can argue with that. While WE08 made some effort to quantify confidence intervals (albeit under very optimistic assumptions), they did not show those confidence intervals in the grading figures, and the figures are what everybody looks at. So in some respects this point is not so much a criticism of WE08 as it is a caveat to the reader of that paper. [However, I do have to say that the grading scale of WE08 does not seem to respect the notion of a confidence interval: a model has a large probability of being 1 sigma away from the estimated truth by chance, but only a small probability of being 3 sigma away: yet the difference between 3 sigma and 2 sigma (0.33) is graded as being equivalent to the difference between 1 sigma and zero (also 0.33).]

It is furthermore clear that since neither models nor measurements are perfect, these errors need to be taken into account in the confidence intervals. And this WE08 did not attempt to do. Moreover I feel that WE08 misspoke when they said on p.5702 that the uncertainty in the observations does not affect the ranking of the models. Surely it must do, if the mean is estimated incorrectly from observations (even for perfect observations, in fact). More importantly, though, it is clearly invalid to assume zero measurement bias and take the interannual variability as the uncertainty in the measurement, because interannual variability is a random error whereas bias is not, and the two behave very different under averaging. Grewe and Sausen rightly distinguish between these two sources of error. I recognize that multiple long-term data sets are not always available for the stratosphere, but every instrument is validated, and the systematic as well as random error is estimated (or at least bounded) from that validation exercise. These estimates should be available, and should be used. Grewe and Sausen apply realistic measurement errors in their simulations and show that these can have a strong effect on the outcome of the grading, significantly widening the confidence intervals. This is a very important point, and I think they have made it well. So, overall, I think this is a useful paper that will be an important contribution to the growing literature on model grading. And I have to agree with the authors that any model grading needs to quote confidence intervals that take into account both model and measurement error. In particular, I agree with their recommendations in the final paragraph of their paper.

Having said that, I do have a major concern with this paper. It basically concludes that the grading in WE08's Figures 2-4 is meaningless. I suspect that this may be a very misleading if not disingenuous conclusion, because they don't actually demonstrate it. What they show, rather, is that a random collection of models with randomly assigned mean errors of up to 3 sigma in individual diagnostics are not statistically distinguishable when considered over 16 different diagnostics. I can believe this result. But is that really the situation considered by WE08? I think not.

Consider WE08's Figure 1. Several models here have mean errors of more than 3 sigma. I just cannot believe that those outliers are correct, and that these differences with observations are not meaningful. Indeed, those modeling groups have subsequently made great efforts to improve their representation of Cly, which would seem to suggest that they believed they had problems. My understanding is that these improvements were generally based on sound physical principles, not just ad hoc adjustments. To me, this is exactly how process-oriented grading should work; it should lead to improved models through an improvement in the representation of key processes.

And consider WE08's Figure 2. There are some localized white spots there, but the eye is drawn to patterns: diagnostics on which many models perform poorly, or models which perform poorly on many diagnostics. While a poor grade on a single diagnostic perhaps could be obtained by chance, the likelihood of a model getting multiple poor grades on related diagnostics by chance must be much, much less. This, again, is where the difference between systematic and random error is critical.

So, in short, the Monte Carlo simulation used to generate Figures 5 and 6 seems to me to be seriously in error as a criticism of WE08's result for CCMVal-1 models (as distinguished from their methodology), since it does not allow for extreme outliers and it does not take account of correlated errors, both of which were arguably present in the CCMVal-1 models, and both of which represented the key scientific outcome of WE08 in terms of model evaluation. In particular, the key statements in their Abstract:

"Monte Carlo simulations show that this method is not leading to statistically significant gradings. Moreover the difference between two models is hardly significant."

are not supported by their analysis. For this reason, I cannot support publication of this paper in its present form.

I can see two possible ways forward. The first would be to restrict the paper to a demonstration that the confidence intervals quoted by WE08 are significantly widened when there is either model or observation error, without casting aspersions on the outcome of the WE08 grading exercise (in terms of identifying outliers). This actually would leave most of the paper intact. The second would be to do what they themselves recommend in the final paragraph of the paper, and re-do the WE08 analysis for the CCMVal-1 models taking account of model and measurement error. If they can do this and show that the actual WE08 results are meaningless, then their claims would be substantiated.

Minor comments:

p.14145, line 16: I don't think you mean "uncertain" here, because a large sigma_obs would lead to a high not a low grade. I think what you mean is "badly estimated".

p.14146, line 11: "that large" should be "so large".

p.14149, line 12 (and Figure 2): What is the value of N here?

p.14151, line 19: I am confused: Figure 1 is for perfect models.

p.14152, line 5: For clarity, it would be good to insert “with these assumptions” after “distinguishable”.

p.14152, line 17: I would suggest saying “overly optimistic” rather than “misleading”, since WE08 were quite explicit about their assumptions.

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