Atmos. Chem. Phys. Discuss., 9, C3830–C3836, 2009
www.atmos-chem-phys-discuss.net/9/C3830/2009/

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

D. Waugh

waugh@jhu.edu

Received and published: 12 August 2009

This comment is written by Darryn Waugh and Veronika Eyring.

The comment by Grewe and Sausen (hereinafter "GS09") examines the statistical significance of the grading calculations performed in our 2008 paper on "Quantitative performance metrics for stratospheric-resolving chemistry-climate models" (Waugh and Eyring, 2008; hereinafter WE08). We are pleased to see that the authors are continuing the conversation on model grading. We agree with them that more analysis is

needed on the statistical robustness of the grading metric used in WE08, and any other grading metric, and also that there needs to be more consideration of confidence levels and uncertainty in observations (e.g. by considering observations from multiple instruments and platforms) in model grading. Furthermore, we think the framework they use is a useful way to examine these issues. However, we object to their conclusions that the method presented in WE08 is not leading to statistical significant grades and that differences between two models is hardly significant. These negative statements about WE08 are not justified or supported by their analysis. We think the unaccounted uncertainty in the observations that are used in GS09 are unrealistically large, and cannot be used to refute our analysis. In addition, there are many cases where the grades in WE08 Figure 2 for two models are statistically different even if based on the GS09 estimates of significance.

1. Statistical significant grading: In numerous places the authors claim that the WE08 analysis does not lead to statistical significant grading, that difference in grades between models is not significant, and that the methodology does not provide the information it was designed for (e.g., in the abstract and in statements on pg 14142, line 6, 7 and 11; pg 14144, line 5; pg 14152, line 17; pg 14155, line 18, and pg 14156, line 12). These allegations against our paper need to be thoroughly justified. However, GS09 do not perform any analysis of the actual diagnostics / grading performed in WE08. They consider synthetic random datasets, and assume large uncertainties in the observations rather than considering the uncertainties in the actual observations used in WE08. For some of the grades in WE08 there may be low statistical significance because of large unaccounted uncertainties, but they have not shown this. More importantly, in order to justify their generalized negative statements against our paper, GS09 would need to show that this is the case for all grades presented in Fig. 2 of WE08. They don't do this, and there are numerous cases where there are significant differences between grades.

One clear example is Cl_y, shown in Fig. 1 of WE08. Here the uncertainty used in the

grading is the uncertainties in the observations, so the fact that the observations are not perfect has been accounted for. Furthermore, there are models that clearly have unrealistic values (e.g., models with peak Cl_y around or less than 1 ppb in polar regions which are in fact more than 5σ from the mean observations) and grades of zero. These model-observation differences are statistically significant and these models are also statistically different from other models that simulate peak Cl_y around 3 ppb (see Review #2 for more discussion on the significance of model-observation differences for Cl_y). Another example is the tape recorder, where again the uncertainties in the observations are used in the grading. This case is discussed in the comment by Strahan et al. Even in cases where the uncertainties in the observations are not included in the grading there are significant model-observation and model-model differences, see point 3 below.

2. Observations: In their calculations GS09 assume there are large uncertainties in the observations used in the grading that are not accounted for in WE08. If there are uncertainties of factors of 2 and 3 times the interannual standard deviation then the significance of the WE08 grading will be low (if this is not accounted for). However, as discussed in the comment by Strahan et al. and the review of Reviewer #2, the uncertainty in most observational datasets used in process-oriented diagnostics is much smaller than this. GS09 give only two examples, one with 50% and another with factor 3, to justify their parameters. Moreover, neither of these observations were used in WE08, and to make statements about WE08 results it is necessary to consider the same observations and diagnostics.

A quick consideration of the observations used in WE08 indicates that the unaccounted uncertainties are much smaller than assumed by GS09. First, consider the analyzed temperatures used for the polar dynamics tests in WE08. We can assess the uncertainty by comparing the different meteorological analyses, as GS09 did to assess the uncertainty in ozone and HCl. For Temp-NH the interannual standard deviation (σ) for the 20 years of ERA40 analyses is 2.9 K, but the difference between means

from UKMO and NCEP and ERA40 is only -0.14 K and 0.6 K, respectively. This corresponds to α less than 5%! For Temp-SH the corresponding values are σ=3.13 K, UKMO-ERA40=-0.45 K, and NCEP-ERA40=1.5 K which corresponds to α = 14% and 48%. So α is larger in the SH than in the NH but still less than 50%. In fact the larger value results from the ERA40 to NCEP bias and, as stated in WE08, the NCEP data have a well-documented bias (of order 1-3 K) in the Antarctic lower stratosphere during winter-spring. So 48% is an overestimate of α for ERA40 in the SH, and using around 15% would actually be more appropriate. Another example where the uncertainties in observations are much smaller than GS09 claim is the U-SP diagnostic. This is based on quantities shown in Fig. 2 of Eyring et al. (2006): the difference between date of transition to easterlies at 20 hPa (where the grading is performed) from the three meteorological analyses shown in this figure is much smaller than the interannual standard deviation, and using the GS09 method this corresponds to $\alpha < 20\%$.

Unaccounted uncertainties of the magnitude of the three examples above will only have a relatively small impact on the values of g required for statistical significance, and the estimates of confidence levels in WE08 would be an underestimate, but reasonable. There are however cases where the uncertainties are larger. One case is the tropical tropopause temperature diagnostic (Temp-Trop). Here the difference between the meteorological analyses is similar to the interannual variability (see Fig. 7a of Eyring et al., 2006). However, this difference is shown in WE08, and we highlight this as a case where the grading is sensitive to the observations used. In this case many of the differences in grades shown in WE08 will have low statistical significance, but there are still cases where the differences are significant (i.e., cases where g<0).

3. Incorrect conclusions in GS09: Even aside from the suitability of the uncertainties in observations assumed by GS09, some of the statements are not justified. In Fig. 6 of WE08 that shows a comparison of the t-statistic with the grading metric g for the diagnostics Temp-NP and Temp-SP as well as for the transport diagnostics using methane (CH_4-EQ and CH_4-SP), there is a huge spread among the models in the metric g and

also the t-statistic. It is clear that there are many models that are significantly different from the observations and pairs of models that are significantly different. Even using the values listed in Table 1 of GS09 (which are based on calculations assuming only 10 years and uncertainties as large as a factor of 2) there are statistically significant differences between grades shown in Fig. 6 of WE08. These are only four diagnostics, but these examples show that the broad generalized negative statements made in the abstract and elsewhere of GS09 about lack of significance of grades in WE08 are not correct.

Also, we do not see the justification for the statement in the conclusions that "If only a 25% or 50% uncertainty with respect to the standard deviation, then the results for the random models presented here suggests that basically none of the models presented in Fig 2 of WE08 differ on a statistical basis". One reason it is hard to see the justification is that GS09 don't quote numbers for significant values of g for α = 25 to 50% percent (they focus on α between 50% and 2 to 3) and they don't say what they think is required for two models to differ (i.e., how many grades need to differ for two models to differ). However, if we again take the numbers in Table 1 of GS09 as representative numbers then $|\Delta g| > 0.5$ is required for two models to differ (using 5% for significance) for a given diagnostic. A quick inspection of Fig. 2 of WE08 shows that there are a very large number of cases where grades from two models for the same diagnostic differ by 0.5 or more. Hence, in contrast to what GS09 claim, there are many cases where the grades for two models are statistically different based on their own estimates of significance.

4. Misleading statements: In the Introduction of GS09 it is stated that "no confidence interval for the grading value is given in WE08" yet WE08 quote values of the grades required for two models to be different or for a model to be different from observations at 5 and 1% levels (e.g., g needs to differ by 0.3 for two models to differ at 5% level when the number of years (N) is 11). Furthermore, in Section 2.2 of GS09 it is stated that a random model with grade 0.77 would be better than a random model with grade

0.46 and that this clearly shows the limitations of the grading methodology as applied in WE08. However, as mentioned above WE08 quote a value 0.3 for difference between grades to be significant at 5% level for N=11. So the difference between models with grades 0.77 and 0.46 is right at this level for significant (and actually not significant if recalculated for N=10), and so is not really clear evidence of the limitation of the WE08 methodology.

5. Constructive analysis: We think that rather than performing and presenting results for random selection of N and α from a range of values it would better to calculate and show the critical values listed in Table 1 of GS09 for specified values of N and α. For example, it would be useful to plot how the minimum $|\Delta g|$ for two models to differ at 5% level increases as α increases, for several different values of N. Such calculations could be used to examine the significance of grades in WE08 (once further estimates of uncertainty in observations used by WE08 are obtained), and could also be used by other scientists applying the same metric to different diagnostics / models.

Although we have issues with many of the statements in GS09 we think the framework they present is useful and should be pursued. If they repeat their analysis for the actual diagnostics used in WE08 together with best estimates of the uncertainties in the various observational datasets then they will be able to make justified comments about the significance of the WE08 grades.

The limitations of the grading used in WE08 are summarized, and the need for further research stated, in the conclusions of our paper. We therefore again strongly support research on this topic and will be happy to see more statistically robust metrics developed and applied to future CCM studies.

References:

Eyring, V., et al. (2006), Assessment of temperature, trace species, and ozone in chemistry-climate model simulations of the recent past, J. Geophys. Res., 111, D22308, doi:10.1029/2006JD007327.

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