Atmos. Chem. Phys. Discuss., 8, S5119–S5124, 2008
www.atmos-chem-phys-discuss.net/8/S5119/2008/

Atmospheric Chemistry and Physics Discussions

Interactive comment on “Quantitative performance metrics for stratospheric-resolving chemistry-climate models” *by* D. W. Waugh and V. Eyring

D. W. Waugh and V. Eyring

Received and published: 21 July 2008

We thank Dr Grewe for his comments. He has raised several issues regarding the statistics of the method used. When we revise the manuscript we will include more discussion of the statistics and when differences in grades are statistically significant. However, we strongly disagree that no conclusions can be drawn from the results presented in this paper or that the models basically do not differ statistically significantly from each other. There are numerous cases where the differences in grades are statistically significant (see below), and conclusions can definitely be made on differences between models ability to reproduce different processes. We also argue that several of the statements that Dr Grewe tries to refute are statements that we never make in the paper.

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

1. Requirement definition.

We do not agree with the proposed list of requirements, and these are certainly not what we had in mind. We also disagree that we even suggested all of these (requirement 1.2 is the exception).

We never say that a model gets a grade 1 if it perfectly simulates reality. Dr Grewe explicitly notes that the statement we make "If g=1 the simulated fields matches the observations" is different from the statement "model represents reality", and this was deliberate. We defined a grade so that g=1 if the model mean matches the observed mean. The metric used satisfies this, and also requirement 1.2. The idea of g=1 for a perfect model comes from Dr Grewe and not from the manuscript. As stated in the manuscript we are using the metric g in this study mainly because the t-statistics which involves the variance in both the observations and models cannot be applied to all our diagnostics as some lack long enough data records for calculation of variance in the observations. To make this point clearer, it is repeated in the Conclusions of the revised manuscript.

We disagree that grade should be within [g-0.1,g+0.1] with an assumed 5% error (requirement 1.3) or difference in grades should be significant if larger than 0.1 (requirement 1.4). Both these requirements are arbitrary, and not implied in the manuscript. Yes we colored the matrix every 0.1 but this does not imply this is the significant level. We never imply that differences of 0.1 are statistically significant. In fact, on page 10880 we use g=2/3 in discussion of statistically significant differences. We make this clearer in the revised manuscript. See more discussion below.

2. Verification

We cannot reproduce all Dr Grewe's statistics (e.g. first series of Monte Carlo) but this is not that important given that he is falsifying statements that only he made and not statements in the paper, i.e. we agree that 1.1, 1.3, and 1.4 do not hold. But as started above we never said they were, and don't think they need to be. The exception is 1.2, which you have not falsified: the statement "if model mean is 1.5 observed standard deviations from observed mean then grade is 0.5" is true for our chosen metric.

If you want to explore the statistical significance of different grades you do not have to perform Monte Carlo calculations, you can use equation 6 in the paper and standard student-t statistics. As we state in the manuscript (page 10880) if $g<2/3$ then there is an extremely high probability that model and data are from different distributions. At the 5% significance level the model and data are significantly different if $g<0.70$ for n=11 years of data (case shown in figure 6b) and $g<0.78$ for n=20 (case in figure 6a). We will include these numbers in the revised manuscript.

As stated above the proposed criteria that grades differing by 0.1 should be significant is totally arbitrary. The above analysis for a specific case of equivalent t statistics suggests that grades need to differ by 0.2 to 0.3 (depending on n) to be significant at the 5% level. Criteria 1.3 and 1.4 are obviously not met, but this does not mean that no conclusions can be drawn from the results or that the models do not differ statistically significantly from each other. The values of g vary from 0 to over 0.9, so there are many cases where the grades differ by much greater than 0.3 and we can make conclusions regarding differences of a model from observations and differences between two models. Differences in the ability of the models to simulate a certain process are very obvious in Figure 1 of this paper and in many if not all Figures of E06. Also, Figure 6 shows an extremely large range in both the g and t statistics.

3. Robustness

Dr Grewe is correct that we have not tested the robustness of the mean model grade to the choice of diagnostics for a larger set of diagnostics than that applied in E06. However, in several places we clearly state that other diagnostics need to be considered and more research is needed to evaluate the key diagnostics and weighting of them. This analysis is only an initial step and it is not possible to look at all possible diagnostics. We have to start somewhere, and the two anonymous reviewers and we feel that

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

focusing on diagnostics from Eyring et al. (2006) is a reasonable starting point.

We discuss in various places that grades will vary if different observations, diagnostic tests, locations evaluated, or metrics are used. But this lies in the nature of the problem. Differences may occur if different observations are used, but we discuss and show examples of this. Also, it is obvious that if observations are very different then the grade will be very different (whatever metric is used), but this applies to all graphical model-data comparisons (shown in numerous other papers). In the revised manuscript we say more explicit that the weighting section is only illustrative and a lot more research is needed (additional diagnostic tests, sensitivity to weighting, etc.) before anything conclusive can be said and a best estimate can be derived.

4. Applicability

We disagree with the statement that no conclusions can be drawn on the quantitative evaluation of the ability of CCMs to simulate future ozone because we have not looked at ozone. In fact, we argue that looking at ozone is not permissible, since this is the quantity of interest. We are focusing on the processes that determine ozone. Yes the list of diagnostics applied is not complete and more work is needed. As stated in various places in the manuscript we have not considered all possible diagnostics and acknowledge that diagnostics on chemistry and radiation should be explored. However, we don't use ozone in the grading as we aim to grade processes that impact on ozone and not the ozone field itself. This work provides a first framework that will enable quantification of model improvements and assignment of relative weights to the model projections that can be built upon.

5. Quality of observational data

Dr Grewe is correct we have not discussed in detail how to include the different types of uncertainties in observations into the grading. However, we present one example where there are large differences in the observations but where we know from other studies that one data set has a bias. In this case the biased data set is not used in the

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

grading. If data sets from different instruments differ and if it cannot be judged whether one data set is better than the other, the key issue will be how to combine these into a single measure of the uncertainty sigma_obs. This is not a straightforward issue, and will need to be carefully considered. We have added a sentence that mentions this and state clearly that observational uncertainties can influence the outcome of model-data consistency tests.

6. Further practical examples

NH Ozone:

Again this discussion is about "perfect" models and not grading the difference between model and observations. As stated above we never make comments about models being perfect. We are quantifying the differences between model and observed mean values.

Ozone Diagnostics:

Again, Dr Grewe is claiming we make statements we don't. We do not say that the model and data have equal variance is a general result. We make this assumption so we can derive a simple analytical relationship between g and t, and then test this for 4 (not 2) cases and show that this relationship is a reasonable approximation. As Dr Grewe has shown there are cases when it may not hold, but this does not mean it is not a useful relationship.

7. Others

Our statements about metrics in Reicher and Kim (2008) and Gleckler et al. (2008) are correct. The "error variance" used in the former is, as we state, the squared difference between model and observed climatological mean divided by the observed variance, and hence similar to the metric used. We state that Gleckler et al. use RMS errors. We see no reason why this will make a big difference.

Interactive comment on Atmos. Chem. Phys. Discuss., 8, 10873, 2008.

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper