Review of Comment on "Quantitative performance metrics for stratospheric-resolving chemistry-climate models" by Waugh and Eyring, by V. Grewe and R. Sausen

This paper (Grewe and Sausen (GS09)) is a commentary on an earlier ACP paper by Waugh and Eyring (WE08)) which proposed a grading scheme for CCMs being used to calculate and project stratospheric ozone abundances. The WE08 study represents a plateau in the multi-year-old CCMVal effort (http://www.pa.op.dlr.de/CCMVal/) to bring some measure of performance metric to the world of CCM models that focus on stratospheric ozone. Without a metric and an implementation framework, scientists are left doing things they don't like to do; namely, giving credibility to predictions from models that they think have demonstrably inadequate skill in one or more key aspects controlling stratospheric ozone. Thus, creating a metric is a good thing in general, as is acknowledged by this reviewer, by GS09, and by authors of other comments on GS09. Differences have arisen, unsurprisingly, in exactly how those grades are derived and represented. The GS09 study is valuable because it expands the framework for considering the precision in the proposed CCMVal intercomparisons.

I have the following comments and specific suggestions that need to be considered by the authors before this manuscript will be suitable for publication.

1) GS09 state (p14156 ln 9-11) '*Hence a statistical(ly) robust grading is absolutely necessary*' to '*better understand model differences and determine specific model shortcomings*.' I respectfully disagree with this statement when it means that we must have it 'now' vs 'eventually' since it is letting the 'perfect be the enemy of the good'. WE08 represents a 'good' approach to this process rather than a 'perfect' process and, hence, is far preferable to the pre-CCMVal years for which there was no evaluation process.

Furthermore, I take issue with the follow-on statement (p14156 ln 14-16): '*In detail, we propose for any future grading (a) to either calculate, estimate, or rely on expert judgment for all of the errors 1–3 described in 2.1, as well as for the inter annual variability;…*' Again this is a description/goal of the perfect world. Simply asserting that one can either calculate, estimate, or rely on expert judgment does necessarily result in the desired outcome. To provide some grounding to this heretical thought, consider the summary diagram of anthropogenic climate forcing from the 2001 IPCC (Climate Change 2001: The Scientific Basis) vs the same diagram from the 2007 IPCC (Climate Change 2007: The Physical Science Basis)) in Figure 1. The 2001 diagram shows best estimates (bars) of climate forcing terms with vertical lines attached. From the 2001 figure caption: '*The vertical line about the rectangular bar with "x" delimiters indicates an estimate of the uncertainty range, guided by the spread in the published values of the forcing and physical understanding*.' and '*The uncertainty range specified here has no statistical basis ….*' In 2007, anthropogenic climate forcing terms appear in a new configuration. This time the now horizontal lines attached to each bar represent '*90% confidence intervals'* following expanded IPCC guidelines for uncertainty expression. In addition, uncertainties were propagated to form an uncertainty for the total anthropogenic forcing for the first time in IPCC assessments. In 2001, with no systematically

quantified uncertainties, the reader could not compare terms reliably, *e.g.*, no one could calculate the probability of the Trop O3 term being larger or smaller than the FF BC term. GS09 might have considered this a total failure on the part of IPCC to have 'no statistical basis' for anthropogenic climate forcing uncertainties in 2001, a topic of some importance in science/policy circles. I assert that the 2001 IPCC authors understood the value of uncertainty analysis but were not able, or not sufficiently comfortable scientifically, to provide 90% confidence intervals. Instead, they provided what they could, a set of numbers (and graph) that was very useful in conveying the qualitative and quantitative message about anthropogenic climate change, and, thereby, set the groundwork for doing more in 2007.

I find distinct parallels between the evolution of the IPCC graphs and the ongoing evolution of the CCMVal grading system. CCMVal is somewhere between the 2001 and 2007 IPCC states of understanding and expression. GS09 has clearly delineated a near-2007 state for CCMVal with a high level of statistical rigor in the grading process. CCMVal isn't that far along as outlined by WE08, but it has already created significant value in the model intercomparison process. GS09 is proposing a standard that would reject this intermediate state for CCMVal. I assert, as have others, that there is considerable value in the current WE08 analysis, which, in addition, sets the stage for 'doing more' in the future. Hence, to close the argument, if the IPCC 2001 authors adopted the rule that '*A statistically robust uncertainty analysis is absolutely necessary to properly quantify climate forcings*.', then the 2001 IPCC figure would not have been published and disseminated at great loss to the climate assessment process. So I suggest that the authors temper their statement in GS09 as '*Hence a statistical(ly) robust grading is* ***highly desirable***' to '*better understand model differences and determine specific model shortcomings*.'

Expanding on this point, I strongly recommend that the authors recast their narrative from one that is predicated on 'absolutely necessary' to one that focuses on 'highly desirable' as note above. This will shift the balance from the requirements of 'perfect' to the value of 'good' as it applies to model grading and shift the balance from 'you must be here (perfection) now' to 'this is how we get from where we are (the good) to where we want to be (perfection).' In other words, describe the path to get, equivalently, from the IPCC 2001 graph to the IPCC 2007 graph. If they recast the narrative as suggested, their colleagues in WE08 and those commenting on this manuscript will be much more accepting of their conclusions.

2) Returning to the statement (p14156 ln 14-16): '*In detail, we propose for any future grading (a) to either calculate, estimate, or rely on expert judgment for all of the errors 1–3 described in 2.1, as well as for the inter annual variability;…*', I think it crucial to note that, while knowing errors is important, it is not always possible (calculate or estimate) or desirable (expert judgment) to derive (p14144 ln10-14):
1 uncertainties in measurement techniques
2. uncertainties in methodology
3. representativeness for a certain region or time
4 representativeness for a climatological value.

In some cases, the uncertainty is simply not known and may be unknowable (*e.g.*, IPCC 2001 graph). Filling gaps with expert judgment can easily become guesswork and set up a *garbage in/garbage out* situation. A good example of unknowable uncertainties comes from considering the age-of-air dataset derived from airborne insitu measurements in the 1990s (see Eyring *et al*., 2006. Figure 10). This dataset stands alone, *i.e*., there are no preceding or succeeding datasets of the same specifications for this crucial stratospheric parameter. Thus, there are no defensible answers to #3 or #4 that are observationally based. On the other hand, most modelers would not forego making a comparison to this unique dataset because of this lack of full statistical information; instead they would use judgment and caveats to qualify their conclusions. As a consequence, in a 'good' grading scheme, use of available geophysical datasets must allow for unspecified uncertainties in order to optimize the value of the few datasets available. So, I suggest that the authors acknowledge the implicit and intrinsic limitations, *i.e*., the reality of actually reaching perfection using the datasets available for intercomparisons.

I also suggest that the authors combine #1 and #2 since the authors' distinction between 'technique' and 'retrieval' here is not very useful. Instead, I can propose 'uncertainties (precision and accuracy) of the reported geophysical quantity(ies).'

3) I strongly support the suggestion of others in commenting on GS09 that the authors redo the WE08 analysis in their (GS09) statistical framework using realistic data uncertainties. Based on the high level of criticism brought forth in GS09, this seems highly warranted and, in prospect, likely to reduce the level of criticism in the revised manuscript. I also hope that it motivates the authors to remove the word 'drastic' from their text.

4) The use of the word 'grade' in this aspect of the CCMVal process is somewhat unfortunate because it implies strong and lasting judgment and hence is likely to elicit emotional responses from readers and participating modelers. However, there is no better word choice, at least in English. It is crucial to acknowledge that the grades as derived by WE08 and GS09 are transient; the default is that they will not survive changes in diagnostics, comparison datasets, or the grading process. Hence, I strongly suggest to GS09, as well as others who create grading results, to always attach a modifier to 'grade' to reduce the 'strong and lasting' implications. Options would be 'current grade', 'phase1 grade', 'test1 grade', etc.

So, with this perspective, we as a community can easily overinvest in the application of statistics to represent the precision of the grading process. What is not well discussed in GS09 or WE08 is the accuracy of the grading process, *i.e*, are we using a set of diagnostics and datasets that are adequate to perform the needed evaluation? An apt analogy is the timepiece that has nanosecond precision but lacks accuracy by minutes. Without inherent accuracy, increasing precision past a certain point no longer increases value. In practice, advances in precision and accuracy are coupled with the IPCC figure evolution being a good example. I strongly suggest that the authors include this thought as a way of establishing perspective, and in some sense a limit, on their drive to increase

precision.

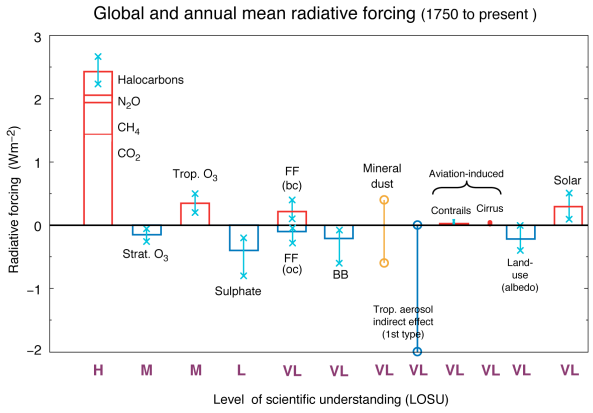
The ultimate limitation in applying grades to models is, and will continue to be, the limited availability of suitable datasets for intercomparison, where suitable is defined by the author's criteria in section 2.1. Limitations in available data will limit the accuracy of the comparison process as discussed above. The authors' more rigorous and comprehensive statistical framework could also be used to identify which datasets are most needed to constrain critical aspects of the model and the required uncertainty for those datasets. This would shorten the path to an accurate grading process and perfect models. I suggest that the authors comment on this.

References

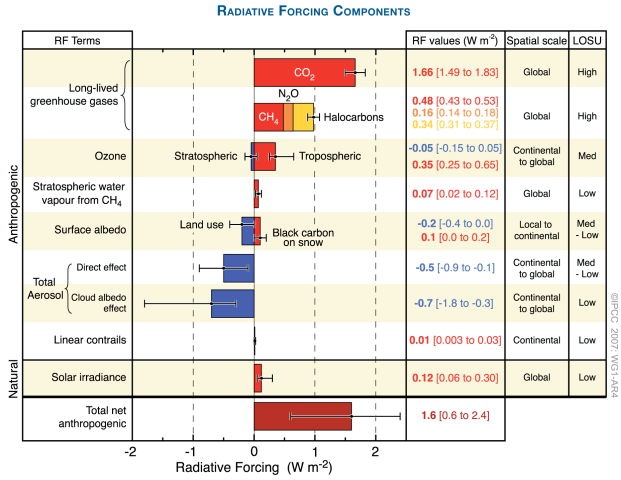
Eyring, V. Assessment of temperature, trace species, and ozone in chemistry climate model simulations of the recent past J. Geophys. Res., 111, D22308, doi:10.1029/2006JD007327, 2006.

IPCC, 2001

IPCC, 2007: Summary for Policymakers. In: *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* [Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M.Tignor and H.L. Miller (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.



IPCC, 2001 (WG1, Chap. 6 p 392, Figure 6.6)



IPCC, 2007 (WG1, Chap. SPM p 4, Figure SPM.2)

Figure 1.