Reply to reviewer #3 on behalf of all authors:

For a better understanding of the reply we cited the comments in black and give our answers in red.

Citing from the referee's comment:

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.

Reply:
We respectfully disagree with the referee. In none of our sentences we have requested a perfect analysis; we agree that "letting the 'perfect be the enemy of the good'" is not the way forward. In our opinion the approach, chosen by WE08, is not a robust approach. And the reason is simply that GS08 applied a statistical method, without analysing the significance for each grading. The result is that the reader of WE08 has no possibility to decide, whether a grades 0.1 or 0.8 are actually good or bad.

Claiming that there were no pre-CCMVal evaluation processes is simply inaccurate and disappointing. A short look at the M&MII report will reveal that most of the diagnostics were already used a decade ago and multi model evaluations were performed. On top of that the same grading as in WE08 was applied 10 years ago (Douglas et al., 1999, Kawa et al., 1999). So in this respect the CCMVal approach is nothing new!
Consequently if this is a first Step, this Step already lasted 10 years!

The matter is not that „Differences have arisen, unsurprisingly, in exactly how those grades are derived and represented", but that grades were calculated which the reader will not be able to interpret. Is 0.3 a good grade or not? From the text (see *Page 5076 in WE08)* it is clear that 0.3 is treated as a bad grade, but, depending on the variability of the model, it might not be distinguishable from reality on a statistical basis.

Citing from the referee's comment:
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.

Reply:
We disagree that this describes a perfect world. A) for some variables WE08 have already estimated uncertainties and used those in the formula. B) WE08 have all data available to actually calculate the inter-annual variability and – without any consideration of an error – to estimate first order confidence intervals.
We respectfully disagree that asking for a statistical analysis when all data are available is a goal of a perfect world.

Further we also cannot agree on the statements concerning the IPCC and uncertainties, exactly for the same argument as above, namely that WE08 used data to calculate mean

values, implies that they could have calculated also standard deviations and confidence intervals just like in GS09. It is not a matter of aiming at a perfect world, but to apply statistics in a reasonable way, which is the more surprising, since the necessary data are available, even if one neglects observational errors. In the IPCC reports climate simulations are investigated and results are only interpreted if a statistical significant signal is detected in order to avoid the interpretation of an arbitrary chaotic behaviour. What we request is to do the same with the grades. In this sense, we see a similarity with the IPCC report.

However, we have re-phrased our main conclusions and avoided words like ‚drastically'.

Citing from the referee's comment:
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.
…
Reply:
Uncertainties in the observations are an important issue. And we think the referee agrees on it. These uncertainties exist. In the case they are not known, this does not imply that this uncertainty does not exist. Not considering it in the calculation, implicitly assumes that this uncertainty is 0, i.e. does not exist. To conclude, the question is what to do with uncertainties, which are not yet estimated: Assume 0% uncertainty (i.e. not considering) or try to find an educated guess.
There is no scientific basis for either answer, but we prefer the second one, since it fosters the discussion and determination of these uncertainties.
We added a paragraph on this discussion in the text (conclusions).

Citing from the referee's comment:
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).'

Reply:
There is a distinct difference between the precision and accuracy of the physical quantity (radiation, column densities) and geophysical quantities, which are often derived from those physical.

Citing from the referee's comment:
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.

Reply:

We agree that it is desirable to calculate robust grades, which are well defined and which give certain information interpretable for the reader. We will be happy to contribute and actually we are planning such kind of investigation. However, we are somehow irritated that this is a critic to our paper rather than to WE08.

Citing from the referee's comment:
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.

Reply:
4.) As stated above the grading formula is at least 10 years old and I currently do not see any significant updates and statistical analysis of this statistical method (the grading is a statistical method since it combined a multitude of information into a single number).

Citing from the referee's comment:
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.

Reply:
We actually deliberately avoided the discussion of the choice of the diagnostics. The evaluation consists of two parts: the diagnostics and the grading (Discussed in detail in the introduction). If the grading is not robust one jeopardises the effort made during the diagnostics, independent from the chosen diagnostic. We had the impression that this issue is not recognized.

Citing from the referee's comment:
… This would shorten the path to an accurate grading process and perfect models….

Reply:
From the final comment, we get the impression that the referee misunderstood the concept of our study. We are not requesting perfect models. We are simply considering what would happen if we had a perfect model, which grade it would get.
This is necessary to get thresholds/confidence-levels, so that one can decide if a model statistically differs from the reality.

References:

Douglass, A.R., Prather, M.J., Hall, T.M., Strahan, S.E., Rasch, P.J., Sparling, L.C., Coy, L., and Rodriguez, J.M.: Choosing meteorological input for the global modeling initiative assessment of high-speed aircraft, J. Geophys. Res., 104, 27545-27564, 1999.

Kawa, S.R., Anderson, J.G., Baughcum, S.L., Brock, C.A., Brune, W.H., Cohen, R.C., Kinnison, D.E., Newman, P.A., Rodriguez, J.M., Stolarski, R.S., Waugh, D., Wofsy, S.C.: Assessment of the Effects of High-Speed Aircraft in the Stratosphere: 1998, NASA-Report, NASA/TP-1999-209237, 1999.