Atmos. Chem. Phys. Discuss., 9, C3837–C3840, 2009
www.atmos-chem-phys-discuss.net/9/C3837/2009/

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

V. Grewe and R. Sausen

volker.grewe@dlr.de

Received and published: 12 August 2009

Reply to the Susan Strahan's, Anne Douglass', and Richard Stolarski's interactive comment by Volker Grewe and Robert Sausen.

We are thankful for the comment given by Susan Strahan, Anne Douglass, and Richard Stolarski. We think that there might be a major misunderstanding of our critics to the paper by Waugh and Eyring (2008) (WE08). In order to clarify this, we like to give 2 statements first:

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

1. There is absolutely no doubt that the existing observational data (in-situ, satellite, etc.) are of immense importance for the understanding of atmospheric processes as well as model evaluation.
2. There is also no doubt that the applied diagnostics, e.g. phase of tape recorder, are of great importance not at least to identify and understand model deficiencies.

If our comment suggests something different, then this was definitely not our intention. Actually, we agree with most of the raised points by Susan Strahan, Anne Douglass, and Richard Stolarski. In order to avoid misunderstandings, we like to clarify what the comment is about: The methodology how to calculate the grades in WE08 is based on a) applying a diagnostic, which includes model and observational data and b) transferring an outcome of this intercomparison into a single number, the grade. Our comment is solely about the second step b).

In the first part of their comment, Strahan et al. raised two issues regarding the volume and accuracy of observational data. These points are more directly addressed in their points 1 to 3, to which we like to answer in the following:

1. We strongly disagree with the conclusion of Strahan et al. that "the abstract (remark: this refers to our abstract in GS09) states that the data used for model evaluation in WE08 assume no error interannual variability". We highly appreciate that most investigations using observational data clearly address uncertainties in the observational data. However, when applying the grading formula, WE08 either included observational errors or an interannual variability in the parameter σ_{obs}. Schoeberl et al. JGR (2008) is actually a good example to investigate the parameter settings for out uncertainty analysis. They calculated an equatorial vertical velocity at 21 km of around 0.025 cm/s $\pm$ 0.008 cm/s (numbers are extracted from their Figure 2). MLS, Recons. H2O and data from Niwando show a discrepancy of around 0.005 cm/s. In this case the bias is around 60%

Full Screen / Esc
Printer-friendly Version
Interactive Discussion
Discussion Paper

($\alpha = 0.6$) of the above mentioned value of 0.008 cm/s, which would fit with our assumptions for the uncertainties. Strahan et al. are totally right that the number of measurements, which are used for all diagnostics are many more than 10 or 40, even thousands and millions of data points. However, that is not the relevant figure/quantity. Relevant are the degrees of freedom, which are much lower, and range between 10 and 20 in WE08. The data used are monthly or seasonal mean values. As an example we focus on the first diagnostic (SON South Polar Temperatures) in WE08, assuming that the data for different years are statistically independent. This diagnostic focuses on a seasonal mean for the time period 1980 to 1999, which gives 20 values used for the diagnostic and for the calculation of the grade. Hence we strongly disagree with the statement of Strahan et al. that the " sampling data have ... insufficient size ... ".

2. New results on CCMVal2 are definitely interesting and useful, and should provide further insights into the grading abilities. However, this has not been addressed in WE08 and moreover, if there is a consistency for different realisations with regard to the grading, then this does not provide any estimate on the robustness and accuracy of the observed mean value. A nice example is given in WE08 (Fig. 5c), where model 7 gets a grade of 0.1 for the UKMO data and a grade of 0.85 for the ECMWF data and this is a significant difference. Even if a number of model simulations get grades around 0.1 for the UKMO based grade calculation, it does not imply that the grade is meaningful if the same calculation based on ECMWF data leads to grades around 0.85. Our point is that a more detailed analysis is needed. Otherwise, how can we be sure that this doesn't happen for other grades? Choosing a range of uncertainties between 0.5 and 3 for the value α is based on two extremes. We assumed that the total ozone timeseries have a small measurement error and HCl measurements a relatively high measurement uncertainty. We would be happy to include smaller values for α, if we see any argument for it.

3. Strahan et al argue that models show the same behaviour for the diagnostics tape recorder amplitude and tropical methane measurements. The grades for 12 models are available in Figure 2 for those two diagnostics. Model 1, 4, and 6 show a difference in the grade of at least 0.6. I.e. 25% of the models do not show this behaviour. In Fig. 7, WE08 show the correlation coefficient of 0.41. The number of degree of freedom is 12 (=number of models). According to e.g. David, F.N., (Tables of the Ordinates and Probability integral ..., London: The Biometrika Office, 1938) the 95% confidence interval of the correlation coefficient is [-0.2,0.75] and hence the value 0.41 does not statistically significantly differ from 0. A statistical significant correlation coefficient has to be at least 0.6 (or -0.6) in order to get a confidence interval [0.05,0.85], which does not include the 0. These high numbers occur only in a few cases in Figure 7 of WE08.

   We agree that there should be a certain correlation between the diagnostics. And we agree that this is probably the case for those mentioned by Strahan et al. However, the conversion into grades seems to include a loss of information, so that as a consequence, most results are not statistically significant.

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