Atmos. Chem. Phys. Discuss., 9, C3417–C3422, 2009
www.atmos-chem-phys-discuss.net/9/C3417/2009/


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

**S.E. Strahan**

susan.e.strahan@nasa.gov

Received and published: 30 July 2009

This comment was written by Susan Strahan, Anne Douglass, and Richard Stolarski.

The main point in Grewe and Sausen (hereafter referred to as GS) is that one should consider sampling issues and the uncertainties inherent in measurements, and how they affect the grading process. GS develop a useful formulation for determining the significance of a particular grade, or the distribution of grades. Unfortunately, they choose parameters for the uncertainty in data that cause them to come to erroneous



conclusions about the usefulness of such data in testing models. These erroneous conclusions, stated in the abstract, are counter to our experience in using observations to grade CCMs with transport diagnostics for the models participating in CCMVal1 and CCMVal2. Although mathematically correct, their statistical example is inappropriate for most diagnostics because it uses observational sample sizes orders of magnitude smaller than many of the actual data sets used for comparisons with models. This paper implies that the past two decades of observations (e.g., T, O3, CH4, N2O, H2O, etc.) measured using satellite instruments (e.g., from LIMS, Nimbus-7, UARS, Aura, SciSat, and Envisat) and aircraft instruments in numerous field campaigns (e.g., from AAOE, AASE, SOLVE, SPURT, Mozaic, etc.) have told us nothing with any statistical certainty about stratospheric composition. The wealth of the literature since the late 1980's, including satellite instrument validations and instrument intercomparisons, argues strongly to the contrary.

Our major objections to this manuscript are described below.

1. There are misleading statements in this paper concerning Waugh and Eyring [2008] (hereafter referred to as WE08). The abstract states that the data sets used for model evaluation in WE08 assume no error or interannual variability. To the contrary, the data sets chosen for model evaluation are selected specifically because analysis of the observations has shown that these data represent some repeatable process at a statistically significant level. Recall that a fundamental idea behind Waugh and Eyring [ACP, 2008] and Douglass et al. [JGR, 1999] is the development of "process-oriented" diagnostics. This often means that the diagnostic is based on a relationship between data in one location or at one time with data at another location or time (i.e., a spatial gradient or seasonal cycle). The uncertainties in such gradients are often very small, as little as a few tens of percent. For example, 15 years of observations are used to derive the phase and amplitude of the water vapor tape recorder in Schoeberl et al. [JGR, 2008]. This data set consists of thousands of water vapor profiles and the means and standard deviations of quantities calculated from such long data sets are

well known. Schoeberl et al. demonstrate remarkable repeatability of this signature.

Instead of the 10 point data set used in the example in GS, let's use the 15-year satellite H2O vapor data set analyzed in Schoeberl et al. as an example. They calculated tropical vertical velocities from these data and found: "The values of the vertical velocity obtained [in this paper are] in agreement with the earlier analysis of Mote et al. [JGR 1996, JGR 1998] using a little over four years of UARS MLS and HALOE data and Niwano et al. [JGR, 2003] using eight years of HALOE data." The tape recorder is not only a repeatable, climatological feature of the atmosphere, but by comparing results from data sets of different length, Schoeberl et al. demonstrated how robust the tape recorder analysis is. The sample data set generated by GS, containing just 10 points, has insufficient size to produce a robust statistical analysis. It is thoroughly misleading for GS to state that their results based on a small data set are equally applicable to all observational data sets used in CCM diagnostics.

Our figure shows the analyzed mean phase of the tape recorder and two standard deviations (black lines) using the Schoeberl et al. analysis [Strahan et al., ACP, 2009]. (CCM analyses courtesy of Dr. Petra Huck, NIWA.) The differences among the models clearly exceed the uncertainty in the phase (ascent rate) in simulations of the tape recorder. While the models falling under the error bars of the observational analysis cannot be distinguished from each other, there is a group of 7 models lying to the left, outside of the error bars, that are clearly distinguishable from the models with good agreement. We conclude it is very sensible to evaluate this fundamental aspect of model performance using a robust quantity such as the tropical tape recorder.

2. From the abstract: "Monte Carlo simulations show that this method is not leading to statistical significant gradings." CCMVal2 Models that have submitted more than one 'realization' of a given reference scenario simulation get very similar grades on a given diagnostic for each of their realizations. The evaluations we perform for CCMVal2 use 10-15 years of model output for grading. Although CCMs have a random component, they also have many important features that are not all random. If they have fast ascent



in the tropical lower stratosphere in one 15-year time series of output, they have fast ascent in all of the realizations. Numerical scores a model receives do vary by a few percent, but no one would consider scores of 62%and 65% to be meaningfully different. We do not ever find that a model gets a high score on one realization (e.g., 80%) but a low score (e.g. 20%) on another. The comparisons used for evaluation are consistent among realizations from the same model because the chosen processes represent climatological features.

In GS, page 14150 lines 10-19, they give quantitative examples that illustrate the problem with their analysis of grades. Consider the case of HCl data that they quote with alpha=3. This means that the uncertainty in the data is 3 sigma of the interannual variability. (We agree that such a large uncertainty suggests that this is not a good diagnostic for discriminating among models.) As mentioned above, observations used to evaluate model process often involve gradients in space or time of a trace species. These uncertainties in such data are often as little as a few tens of percent. This would correspond to alpha of 0.1 or 0.2, not 3, and would produce a far different GS Figure 2 – nearly all the points would have values above 0.8. In this case, the opposite conclusion would then be reached: the grading process is indeed robust. Had they used a more realistic value for alpha, they would have strengthened the grading process by demonstrating that, objectively, we can say how significant a difference in grades is. Properly applied, the formalism could also be used to help evaluate the usefulness of various diagnostics.

3. From the abstract: "Moreover, the difference between two models is hardly significant." In many cases, the differences between models' grades are significant. We use many diagnostics, and some of them evaluate essentially the same processes but do so using different data sets. If the authors' statement that the difference between models' grades is insignificant were true, then we would not find consistency between the grades given by different diagnostics. The change in tape recorder phase with height gives a measure of the tropical ascent rate. But the difference between midlatitude

mean age and tropical mean age also is a measure of ascent rate. The data sets used for these comparisons are independent, i.e., the mean ages come from CO2 aircraft and balloon data while the tape recorder comes from satellite H2O and CH4 measurements. When applied to a given model, these diagnostics generally give the same information, an unlikely result if the comparisons are meaningless. The same can be said for other pairs of diagnostics (e.g., the tape recorder amplitude and tropical CH4 profile gradients – they both gauge the strength of mixing across the subtropics).

Further, we find a strong correlation between model scores in the tropics and in the polar region, even though in each region completely different data sets are used. Tropical and polar grades are correlated because models that do a good job of getting air into the stratosphere with reasonable ascent and mixing are more like to realistically represent important physical processes (i.e., radiation and wave activity). If a model performs poorly in the tropics it stands almost no chance of having a realistic polar composition. If grades were meaningless, we would not see any correlations.

We do not believe that the 'grading metric' as defined in WE08 is the only possible approach towards objective evaluation of models. WE08 (along with other papers) attempts to open the conversation: how do we directly use observations to decrease the uncertainty in prediction? An example of a different approach from WE08 would be to use the results of objective model evaluation to explain variance in prediction. Unfortunately, GS takes the extreme view that there is no merit to objective evaluation of models as it is currently practiced. We strongly object to the conclusions as written based on our experience with both the observational data sets and the application of diagnostics to CTMs and CCMs. The misapplication of the statistical formalism as demonstrated in GS leads effectively discredits the observations and leads to the erroneous conclusion that model evaluations are meaningless.

---

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



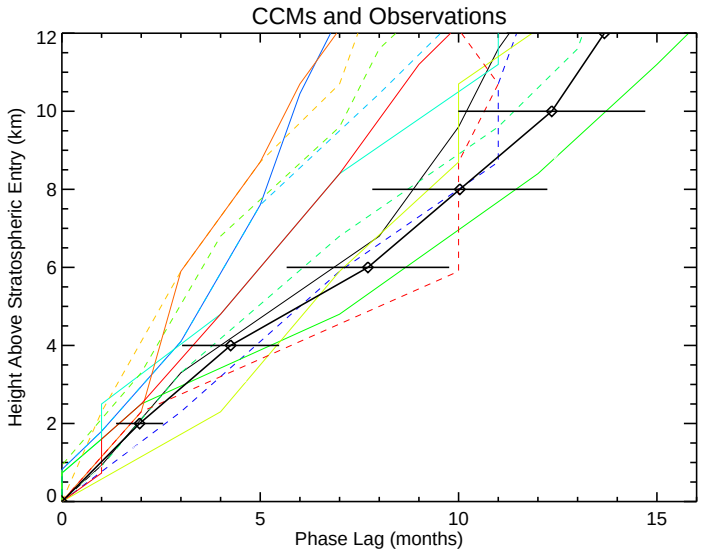


**Fig. 1.**