Atmos. Chem. Phys. Discuss., 8, S4732–S4736, 2008
www.atmos-chem-phys-discuss.net/8/S4732/2008/

Atmospheric Chemistry and Physics Discussions

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

Anonymous Referee #1

Received and published: 12 July 2008

This paper considers the application of quantitative performance metrics to stratosphere-resolving chemistry-climate models. It has to be recognized at the outset that the subject of quantitative performance metrics for climate models is emotionally charged, with some modellers objecting to them almost as a matter of principle. Therefore it is essential to distinguish between criticisms of any sort of metric-based evaluation, and an assessment of this paper in particular.

I discuss first the general issue. Note that there is a distinction between: (i) assessing model performance in a quantitative way, process by process; (ii) coming up with an overall model "grade"; and (iii) weighting consensus projections according to what is relevant for the quantity in question (which is probably the hardest part).

In principle, one could perform (i) without ever moving on to (ii) or (iii). But of course the main driver for developing climate models in the first place is to make projections of future changes, and decision-makers are increasingly asking for climate projections WITH UNCERTAINTIES. The responsibility is then on us scientists to provide the best possible information. The current practice is usually to use the spread of model projections (either standard deviation, or sometimes even range) as an estimate of uncertainty. However there is probably zero scientific basis for this. It is instructive to recall that in Chapter 6 of the 2006 Ozone Assessment, the chapter authors (notably the two authors of this paper) realized that they simply could not use the entire model spread as an estimate of uncertainty of global ozone recovery dates, if they wanted to provide the best possible scientific information to policy-makers, because the outliers (on both sides) resulted from models that exhibited unphysical behaviour for global Cly. So in this case the issue of metrics could not be avoided, as to do so would have been irresponsible in terms of providing the best scientific information for the policy-makers. That is what led to the solid and dashed curves (a simple form of metrics). I know that all may not agree with the way this was done, but my point here is simply to emphasize that the scientific community is obligated to tackle this question of metrics in order to provide the best quality scientific information.

I think that if we look at this as a question of how to quantify scientific uncertainty, rather than model grading, it will become less emotionally charged and more palatable. Nobody would imagine releasing observational data without error bars (and we don't generally view error bars in that context as a "beauty contest"), yet we do so routinely with models. (Well, we quantify sampling error which we assume to be random, but never systematic error.) Perhaps one might argue that's because it can be done for measurements but not for models. But there's a lot of uncertainty in measurement error bars too (especially for the systematic errors), and just because something is difficult doesn't mean one shouldn't try. In fact, we (speaking for the modelling community

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

broadly) already provide uncertainties (though the model spread), so the question is rather whether we can do better than that.

I sometimes hear it said, as a criticism of quantitative metrics, that there is no proven relation between the ability of a model to simulate present-day climate and future changes. I suppose that is literally true because we have no way to assess the ability of a model to predict changes that have yet to occur. But this is quite a remarkable objection, as it completely rejects the accepted standard for building up one's confidence in a model. Imagine trying to publish a paper with future projections if the model had never been compared with measurements in any way! So such an objection simply cannot be taken seriously. The question is, rather, what is the best way to assess one's confidence in a climate model? The entire premise of physically-based (as opposed to statistical) modelling is that if one correctly solves the governing equations, then the correct behaviour of the system will result. From that perspective it seems like a very reasonable approach to compare a model with observations in a process-oriented fashion, rather than by looking at the outputs (which seems to be the approach with the IPCC models). And that is, of course, the whole premise behind CCMVal.

So I don't think there is any escape from the need to assess models by comparing them with measurements, nor is there any escape from the need to provide the best quality scientific information for policy-makers, which may mean de-weighting model projections that are felt to be unphysical or unreliable. Metrics offer the advantage of being able to include all model projections in the ensemble, rather than having to exclude some of them, which among other things is politically more palatable. Of course it's the case that metrics could be used in inappropriate ways. But so can model projections themselves. So it is important that any criticisms of metrics be specific and testable, and not just speculative. Frankly I suspect that many objections to metrics are made because people don't want to see their own model singled out as being below the mark.

Viewed from this context, I find the Waugh & Eyring paper to be a very interesting and important contribution to the emerging literature on quantitative performance metrics for climate models. It is novel in its use of process-oriented diagnostics, it investigates issues such as robustness, statistical significance, and relationships between diagnostics, and it has the strength of being directly traceable to the Eyring et al. papers (2006, 2007) so that readers can easily look at each comparison graphically. So it provides a very valuable basis for the future investigations that the authors readily admit are required. I recommend publication after the following comments are addressed.

Major comments

1. As the authors note, the models generally do very poorly on the polar CH4 diagnostic. But unlike in the case of tropical tropopause temperature and stratospheric entry value water vapour (the other two diagnostics with very poor mean model grades), where the models show a very large spread, in this case Figure 5(c) of E06 shows that the models give very consistent vertical gradients of CH4, it's just that they're considerable stronger than the HALOE gradient. Especially given the limited HALOE coverage at high latitudes, this raises the issue of how reliable the measurement estimate is. This is an important issue to resolve because the authors are effectively claiming that there is a systematic model bias in this respect. Before everybody invests a lot of time looking into this the authors need to confirm their result by comparison with some other data set.

2. In several places (including the abstract) the authors note that the weighted multi-model mean ozone projection is not very different from the unweighted projection. That is not particularly surprising, if the poorer-performing models can have biases of either sign. However what is surprising to me is that the uncertainty seems not to be much reduced by the weighting! That would seem to be much the more interesting result, yet it is not mentioned at all by the authors. This needs to be remedied (and discussed).

3. I understand that Figure 8 is the way it is for traceability with E07. But it has to be recognized that the (narrow) uncertainty range on the observations is really misleading,

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

because of the uncertainty in defining the baseline. From Figure 8 one would conclude that CCMs are systematically underpredicting NH midlatitude column ozone depletion by something like a factor of two. Shepherd (2008 Atmos-Ocean) showed that when the baseline is defined by 1965-1975, the NH midlatitude ozone depletion in CMAM agrees extremely well with the observations, even though from this figure one would conclude it is off by a factor of two. This difference has to do with trying to define a baseline from post-1980 values alone. I appreciate that this doesn’t affect any of the conclusions that are being drawn here, but this problem should at least be acknowledged so that people don’t draw the wrong inferences from this figure.

Minor comments and typos

4. I’m a little confused by equation (3). If one substitutes g_k = 1 then one gets N/(N-1) times the multi-model variance. But shouldn’t we be getting something more related to the standard error of the mean, which decreases with increasing N?

5. p.10879, lines 19-20: I am confused by this statement since the variance in the observations is a parameter in equation (4). Or does sigma_obs mean something different in equations (4) and (5)?

6. p.10880, line 15: For clarity, I would insert "overall" after "the models".

7. p.10888, line 23: Delete "is" after "grades".

8. p.10890, line 2: "meteorological" is misspelled.

9. p.10890, line 5: "O’Neill" is missing an apostrophe.

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

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper