Atmos. Chem. Phys. Discuss., 8, S3695–S3703, 2008
www.atmos-chem-phys-discuss.net/8/S3695/2008/

Atmospheric Chemistry and Physics Discussions

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

V. Grewe

volker.grewe@dlr.de

Received and published: 13 June 2008

Interactive comment on the manuscript 'Quantitative Performance Metrics for Stratospheric-Resolving Chemistry-Climate Models' by Darryn Waugh and Veronika Eyring.

Volker Grewe, Deutsches Zentrum für Luft- und Raumfahrt, Institut für Physik der Atmosphäre, Oberpfaffenhofen, Germany

The authors aim at grading models with respect to their ability to predict future ozone changes.

In general, I do appreciate the authors' effort to establish quantitative performance metrics. The authors' approach for a model validation based on a multi-model basis is well chosen, since it has the potential to offer many insights, e.g. it might indicate areas, which generally need to be better understood.

However, in my view, the chosen grading approach insufficiently summarises the abilities and disabilities of chemistry-climate models. I have strong objections that any conclusions can be drawn from the results, e.g., from Figs. 2-5.

My main concern is that the statistics of the method is not sufficiently investigated. A requirement definition and a verification of the grading are missing. This would imply a list of requirements such a grading should comply with. This should be completed by a verification, i.e. a proof that the proposed method/grading actually complies with these requirements.

This is an important issue, since without, the significance and implications of the results are not known.

Alternatively, instead of giving a requirement definition and a verification, a sophisticated analysis and interpretation of the grading g could be given, adressing questions like: "What grading gets a perfect model?"; "When do two model gradings differ significantly?", etc.

Further, a robustness of the mean model grading would be necessary. Finally, I question the applicability of the method to evaluate the models' capability to simulate future ozone.

All these issues are discussed in more detail below.

I hope that my comments stimulate the authors to include a more sophisticated statistical basis, which probably will lead to the answer that the models basically do not differ statistically significantly from each other, or to revise the methodology significantly (including a profound testing of the method).

ACPD

8, S3695–S3703, 2008

Interactive Comment

1. Requirement definition The grading aimes at comparing a mean value from a certain diagnostic with a mean value from observations. The differences are scaled by three times the observed inter-annual variability (in terms of standard deviation).

From that I would deduce the list of requirements (1.1 to 1.4) for the grading, which can be found below. Note that this simply reflects what I think the authors could have had in mind when defining this grading. The first two points (1.1 and 1.2) are not explicitely mentioned in the text, but somehow suggested in section 2.2 (2nd sentence). Note that the sentences (1.1) and the respective sentence in the manuscript (10878, bottom) differ, though they sound similar ('g=1 $\Rightarrow$ model matches observations' versus 'model represents reality $\Rightarrow$ g=1'). It might sound picky, but I think it is essential to the grading. The last two (1.3 and 1.4) address the question, when do two grading differ significantly? This is not at all addressed in the manuscript, which is one of the major point of critics I have. The statements 1.3 and 1.4 reflect what I think would be necessary for a model grading. Anyway, a less stringent definition of the requirements 1.3 and 1.4 may be acceptable, but still they have to be proven. In section 2.2 I will show that two model gradings are basically not distinguishable.

Grading characteristics / requirements, which are not fullfilled:

1.1 A model gets the grade 1, if it perfectly simulates reality.

1.2 A model gets the grade 0.5 if the model's mean value differs by 1.5 times the standard deviation of the observation.

Further, I would expect a grading to give results which are statistical significant within a certain range. This leads to 2 further requirements:

1.3 The confidence interval of g should be [g-0.1,g+0.1] with an assumed error of 5% (e.g.), or the real 'g' should be greater than the estimated g minus 0.2 (one-side confidence interval).

ACPD

8, S3695–S3703, 2008

Interactive Comment

1.4 The difference of the gradings of 2 models should be significant, if it is larger than 0.1, with an assumed error of 5%.

Explanations:
1.3 gives an indication on the quality of the obtained grading value. How certain are we that the color coding in Figure 2 is adequate? If the confidence interval is ± 0.1 then this would basically result in 5 differing colors. Any finer color coding would imply a degree of accuracy, which is not given. If the confidence interval is ± 0.3, then basically only 2 colors would correctly reflect the degree of accuracy.
1.4 just decribes how well the grading of two models can be distinguished. If this can only be done within a large uncertainty, then the weighting of the model results to get a mean model result (see eq. 2) is meaningless.

Note that the assumed error of 5% implies that the grading is out of the confidence interval in around 10 cases in Figure 2. In other words, even with a color coding, which respresents the accuracy of the grading g, in 10 cases the assigned grades (=colors) have to be expected to be wrong, anyway.

2. Verification 1. A perfect model and perfect observations.

Let us assume we have a perfect model and we know the reality of the regarded diagnostic perfectly and that the values of the individual years (here: 10 years) are distributed normally. I.e. the reality can be represented by the expectation μ and a standard deviation σ. Since model and observations are considered to be perfect in this case, the model's and observational's expectations and standard deviations are equal: $\mu_{mod} = \mu_{obs} = \mu$ and $\sigma_{mod} = \sigma_{obs} = \sigma$.

What is the expectation of the grading? Is it simply: $\hat{g} = 1 - \frac{1}{3}\frac{\mu - \mu}{\sigma} = 1$? No, because every individual realisation of the 10 year period gives a mean value, which differs from μ.

This can be calculated by using a Monte Carlo simuation (here: 100,000

ACPD
8, S3695–S3703, 2008

iterations). The derived mean value for g is 0.87 (Median 0.88) and hence differs remarkably from $\hat{g} = 1$. Note that the conclusions are the same if equally distributed values are assumed instead of normally distributed values.

Hence, one cannot verify point 1.1 from the list of requirements, i.e. 1.1 is falsified.

Calculating the 95% (99%) percentile from the frequency distribution gives a value of g of 0.65 (0.5) for this case, i.e. a perfect model and perfect observations!

Hence, in the case of an assumed error of 1% a value between 0.5 and 1 will be derived, which disagrees with statement 1.2. Hence 1.2 is falsified.

2. A perfect model and imperfect observations.

   We know that observations have errors from measurement techniques, from analysis and due to spatial sampling or certain conditions under which the observations are derived. They also have some uncertainties related to the representativity (e.g. Lary and Aulov, JGR 113, 2008).

   Let us assume that (as above) the realitiy can be described by an expectation μ and a standard deviation σ. The expectation μ_{obs} and standard deviation σ_{obs} of the observations are imperfect and differ from the real values μ and σ. Let us express these differences in relation to the standard deviation σ. With an error α for the expectation and and error β for the standard deviation one gets $\mu_{obs}(\alpha) = \mu + \alpha \times \sigma$ and $\sigma_{obs}(\beta) = (1 + \beta) \times \sigma$. If we now look for the 95% and 99% one-sided confidence interval, we get the following results of g, if we consider $\pm$10% and $\pm$50% variations for errors α and β:

ACPD

8, S3695–S3703, 2008

	95%				99%			
$\alpha \backslash ^{\beta}$	-50%	-10%	+10%	+50%	-50%	-10%	+10%	+50%
-50%	0.62	0.66	0.64	0.62	0.48	0.54	0.50	0.48
-10%	0.52	0.70	0.70	0.68	0.36	0.58	0.58	0.56
+10%	0.44	0.68	0.70	0.70	0.22	0.58	0.58	0.58
+50%	0.26	0.60	0.68	0.72	0.01	0.54	0.54	0.62

In the case of 10% uncertainties in the observations, model gradings larger than ≈0.7 (5% error) and ≈0.6 (1% error) have to be regarded to be perfect. However a 10% uncertainty in the observations is rarely obtained. Let us assume that the observational variability is underestimated by 50% and the mean value overestimated by 50% of the standard deviation. Then all models with a grading, which is larger than 0.26 and 0.01, have to be assumed to be perfect with an assumed error of 5% and 1%, respectively. This clearly contradicts statements 1.1, 1.2, and 1.3.

Note that the authors already discussed partly some of these issues (page 10880), when they concluded that a significant difference can be found for g<2/3 (assuming perfect observations, though). However, the implication for the Figures 2-4 are not discussed. E.g. some of the models might be regarded to be indistinguishable from a perfect model for some of the diagnostics. What implications does it have for the mean model in Fig. 3?

3. 2 identical, but imperfect models and perfect observations

   Here the difference of two model gradings is investigated. Let us assume that we have perfect observations and two identical models. I.e. the reality can be expressed with μ and σ. For the perfect observations we have therefore μ=μ_{obs} and σ=σ_{obs} The two models are equally imperfect with an expectation $\mu_{mod}(\alpha) = \mu+\alpha\times\sigma$ and standard deviation $\sigma_{mod}(\beta) = (1+\beta)\times\sigma$, as above.

   The expectation of either model is $\hat{g}$ and the difference of both gradings is $\hat{g}$-$\hat{g}$=0.

ACPD

8, S3695–S3703, 2008

To estimate the uncertainty ranges, I performed a Monte-Carlo simulation for $\alpha = \beta$ (arbitrary choice) in the range of $[0, 0.5]$, which are indeed small values for current CCMs. The mean 95% and 99% percentiles of the difference of the gradings of the two models are 0.48 and 0.57, respectively. However, 5% of the differences are larger than 0.85 and 1% are larger than 0.9.

This leads to the conclusion that two gradings are basically not distinguisable, unless the difference is larger than 0.9.

This falsifies statement 1.4.

To summarise this part: If the requirements of the grading are as specified in 1.1 to 1.4 then the calculations above show that the grading fails. Or to formulate it more positively: the given grading implies for a perfect model and perfect observations a mean grading of 0.87. If observations have an error of 10% error, then models with a grading value larger than 0.58 have to regarded to be perfect. There is a 5% chance that the grading of two identical models differ by more than 0.85.

3. Robustness So far, the gradings for individual diagnostics were investigated in more detail. The authors deduce from those a mean model grade. However, a robustness of this grading with respect to the chosen diagnostics is not given. This could e.g. include an indicator for the variability of the grading for a larger number of subsets of the diagnostics.

A nice example how such a robustness can be derived is shown in Reichler and Kim (2008), who investigated the 95% precentile for a large number of subsets.

4. Applicability The aim of the investigation is to perform a quantitative evaluation of the ability of CCMs to reproduce key processes for stratospheric ozone.

In fact, none of the diagnostics include the model's ability to simulate ozone, past ozone trends, ozone variability pattern, etc., nore any diagnostic concerning radiation, photolysis, clouds, or other climate issues. In principle, a good transport

ACPD

8, S3695–S3703, 2008

model without any reasonable ozone chemistry has the potential to be graded with high values. The ozone forecast skills, however, of such a model, are extremely low. Hence from the grading as it is perfomed, no conclusions can be drawn on the quantitative evaluation of the ability of CCMs to simulate future ozone.

5. Quality of observational data The authors acknowledge that there may be some uncertainties in the observational data. However, it is not shown how these may be considerered in the grading. I.e. if inter-annual variability, measurement errors, analysis uncertainties, and uncertainties due to representativity or sampling are known for certain data, how should those be combined for the grading?

6. Further practical examples 1. NH ozone

Assume that a diagnostic considers NH winter ozone values. In the observations a ten year period is taken into account, which has no major warmings. In the model, in the first and last year a major warmings occur, which leads to an increase of ozone by 40%. (E.g. Obs.: 5.1, 5.1, 5.2, 5.1, 4.8, 4.7, 5.0, 4.9, 4.9, 5.1 ppmv). Hence the mean value and the variability increases from around 5 to 5.3 ppmv and 0.16 to 0.66, respectively. The grading is 0.06, however, the mean value does not differ statistically signifcant.

In that sense, the model can be regarded to be perfect, however the grading gives a value of almost 0.

2. Ozone diagnostics from Eyring et al. (2006)

In the text the authors give an example for statistics, where they assume that the standard variabilities of the model and observational data are equal. What reason exists to assume that both standard variabilities are equal? Why should that hold on a general basis, even if it holds for 2 cases presented in 4.2 (T and CH_4)? As an example I show below standard variabilities presented in Eyring et al. (2006) (Tab. 3, Antarctic ozone depletion):

ACPD
8, S3695–S3703, 2008

Diagnostic	Obs	Model range	Factor difference
Min-SH	11.2 DU	4.4-38.2	3
OHA	2.2 mio km^2	1.6-5	2.5
OMDobs	5.4 mio tons	1.8-9.6	2

The standard variabilities differ by up to a factor of three! To illustrate it more, one can take, as an example, the values from CMAM und NIWA. Mean values do not differ statistically for Min-SH. (t-Test value 1.6 < 2.1 with assumed error of 5%). The grading is, however, relatively low, with 0.36.

7. Others Section 2.2. (page 10878, lines 16ff) To avoid misunderstandings: As far as I understand, Reichler and Kim (2008) based their method on error variances, not mean values. Gleckler et al. (2008) applied RMS errors to account for spatial and pattern and the annual cycle. This could well make a big difference and should be made clearer in the text.

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