Atmos. Chem. Phys. Discuss., 8, S4343–S4344, 2008
www.atmos-chem-phys-discuss.net/8/S4343/2008/

Atmospheric Chemistry and Physics Discussions

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

P. Braesicke

peter.braesicke@atm.ch.cam.ac.uk

Received and published: 1 July 2008

The Waugh and Eyring paper is certainly starting a thought provoking discussion and I think we are in danger mixing-up two very distinct questions about grading:

1) When we use an established diagnostic (or develop a new one) which allows us to compare models to observations, it is certainly fair to point out that a particular model reproduces a certain aspect of a selected diagnostic better than another model, and under certain circumstances it will be possible to quantify the quality of the agreement/disagreement between models and observations (see the caveats raised in Volker Grewe's comments). This might be used to rank models for a particular diag-

nostic only. This is an active part of research in which boundaries between research areas are crossed (modelling, remote sensing, in-situ observations, data assimilation, operational services, climate centres, etc.), and which corresponds to individual rows in Figure 2 of Waugh and Eyring.

2) In contrast, if we are looking into an abstract question which in itself is beyond direct verification ("When will ozone recover to 1980 values?", "What will the ozone be in 2100?"), it seems dangerous to pre-judge any outcome of model simulations by defining a necessarily limited set of diagnostics and introducing a rather arbitrary weighting to each diagnostic. Two problems seem to arise here: What works for the past might not work for the future (physically and in terms of model parameterisations) and if modellers know the "goalposts" they will inadvertently try to "score" and "nudge" their models into the "right" direction (because in a fair assessment a set of diagnostics will be defined prior to the assessment). Just to clarify: Having e.g. a report on the state-of-the-art in chemistry-climate modelling is certainly a desirable aim, to come up with an ultimate grade for each participating model (equivalent to Figure 4) might produce a wrong sense of confidence in a particular model.

The authors' own discussion mentions that their weighted result is similar to the "classically" averaged result. I believe that there is a lot to explore in 1); I do not believe that there is a benefit of a unified rating system which issues a single grade to a model, because even a simple, weak performing model might be extremely useful when applied properly and will not show up in a multi-model average.

Finally, introducing single grades for a pre-defined subset of diagnostics will only introduce a wrong sense of confidence in policy makers, who will start choosing results from particular models because of their particular grading (bypassing expert knowledge) - a situation that should be avoided. I am looking forward to a discussion in the community about how to proceed and how to address the obvious caveats of this approach.

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

ACPD

8, S4343–S4344, 2008

Interactive Comment