Quantitative Performance Metrics for Stratospheric-Resolving Chemistry-Climate Models

Darryn W. Waugh[1], and Veronika Eyring[2]

[1]Johns Hopkins University, Baltimore, USA

[2]DLR Institute of Atmospheric Physics, Oberpfaffenhofen, Germany

Introduction

A set of performance metrics is applied to stratospheric-resolving chemistry-climate models (CCMs) to quantify their ability to reproduce key processes relevant for stratospheric ozone.

Assigning quantitative metrics ("grades") to a range of diagnostics will

- Allow easy visualization of model performance
- Identify systematic biases and falsely modeled processes
- Enable quantitative assessment of model improvements
- Enable weighting of model predictions.

Grading Metric

The simple metric (Douglass et al. (1999)

$$g=1-\frac{1}{n_g}\frac{|\mu_{\rm model}-\mu_{\rm obs}|}{\sigma_{\rm obs}}$$

is used for all diagnostic tests, where μ_{model} is the climatological mean from the model, μ_{obs} is the observed climatological mean, σ_{obs} is the uncertainty in the observed mean, and n_g is a scaling factor (=3, here). If g=1 the mean from the model matches the observed mean, while if g=0 the mean is 3σ from the observed mean.

This metric can be applied to all diagnostics, is easy to interpret, can be easily compared between tests, and can be related to the student t-test (see below).

Models and Diagnostics

We consider the CCM simulations, diagnostics, and observations that were shown in Eyring et al. (2006).

The 13 models are listed in Table 1.

The diagnostic tests applied to simulations of late 20th century are listed in Table 2. Each diagnostic is based on a comparison shown in Eyring et al. (2006).

Name	Reference
AMTRAC	Austin et al. (2006)
CCSRNIES	Akiyoshi et al. (2004)
CMAM	Fomichev et al. (2007)
E39C	Dameris et al. (2005)
GEOSCCM	Pawson et al. (2008)
LMDZrepro	Lott et al. (2005)
MAECHAM4CHEM	Steil et al. (2003)
MRI	Shibata and Deushi (2005)
SOCOL	Egorova et al. (2005)
ULAQ	Pitari et al. (2002)
UMETRAC	Austin (2002)
UMSLIMCAT	Tian and Chipperfield (2005)
WACCM	Garcia et al. (2007)

Table 1. CCMs used in this study.

Short Name	Diagnostic	Quantity	Observations	Fig. E06
Temp-NP	North Polar Temperatures	DJF, 60°-90°N, 30-50hPa	ERA-40	1
Temp-SP	South Polar Temperature	SON, 60°-90°S, 30-50hPa		
U-SP	Transition to Easterlies	U, 20hPa, 60°S	ERA-40	2
HFlux-NH	NH Eddy Heat Flux	JF, 40°-80°N, 100hPa,	ERA-40	3
HFlux-SH	SH Eddy Heat Flux	JA, 40°-80°S, 100hPa		
Temp-Trop	Tropical Tropopause Temp.	T, 100hPa, EQ	ERA-40	7a
H_2O-Trop	Entry Water Vapor	H_2O, 100hPa, EQ	HALOE	7b
CH_4-EQ	Tropical Transport	CH_4, 30/50hPa, 10°S-10°N, March	HALOE	5
CH_4-SP	Polar Transport	CH_4, 30/50hPa, 80°S, October	HALOE	5
CH_4-Subt	Subtropical Tracer Gradients	CH_4, 50hPa, 0-30°N/S, Mar/Oct	HALOE	5
Tape-c	H_2O Tape Recorder Phase Speed	Phase Speed c	HALOE	9
Tape-R	H_2O Tape Recorder Amplitude	Amplitude Attenuation R	HALOE	9
Age-10hPa	Lower Stratospheric Age	50hPa, 10°S-10°N and 35°-55°N	ER2 CO_2	10
Age-50hPa	Middle Stratospheric Age	10hPa, 10°S-10°N and 35°-55°N	CO_2 and SF_6	10
Cl_y-SP	Polar Cl_y	80°S, 50hPa, October	UARS HCl	12
Cl_y-Mid	Mid-latitude Cl_y	30°-60°N, 50hPa, Annual mean	multiple	-

Table 2. Diagnostic tests used in this study.

Example: Inorganic Chlorine

As an example, consider first the grades for the Cl_y tests. Fig. 1 shows time of mid-latitude and polar Cl_y from the models and observations. As noted in Eyring et al. (2006), there is a large spread in modeled Cl_y with some large mode-data differences. For polar Cl_y, some models (1,11, and 12) produce values close to the observations and have g>0.9. In contrast, several are more tan 3s from the observations and have g=0.

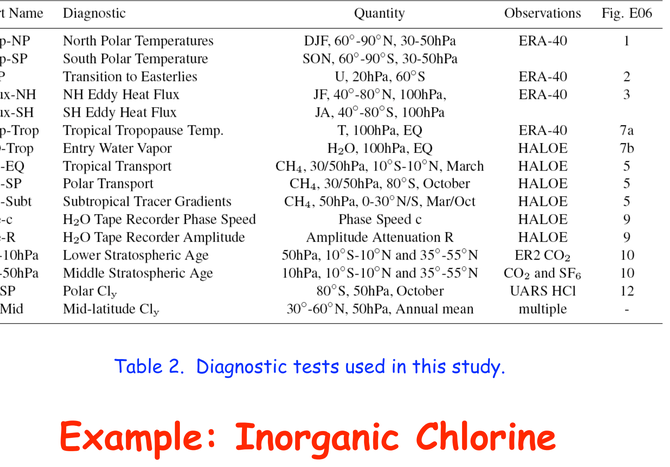

Figure 1. Times series of 50 hPa Cl_y for (a) annual mean 35-60N, and (b) October-mean 80S from CCMs (curves) and observations (symbols). Lower panel shows grades for these two tests.

Model Grades

The grades for the application of each test to each model is shown in the "portrait" diagram in Fig. 2. There is a wide range of grades (g=0 to 0.9), with large variation for different diagnostics applied to same model (columns), and for same diagnostic applied to different models (rows)

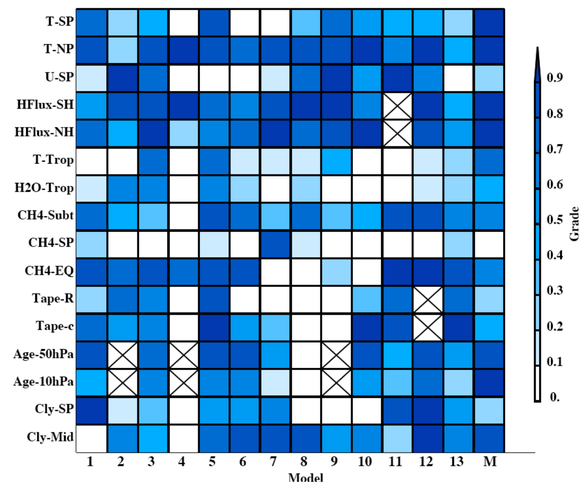

Figure 2. Matrix displaying grades for application of each diagnostic to each CCM.

There are no tests where all model perform well or all models perform poorly. However the majority of the models perform well for NP temperature and the heat flux tests, whereas majority perform poorly for tropopause temperature, entry water, and polar CH_4, see Fig. 3.

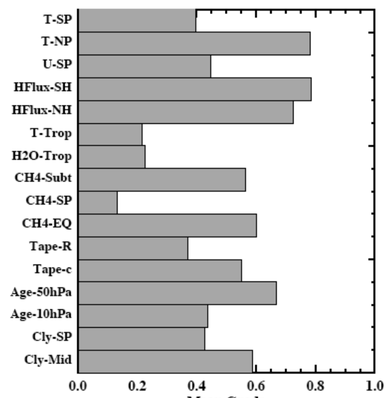

Figure 3. Average grade for each model.

No models score high or low on all tests, however differences in the performance of the models can be seen and quantified, see Fig. 4. For example several models get low grades on multiple transport tests and have very low average transport grades.

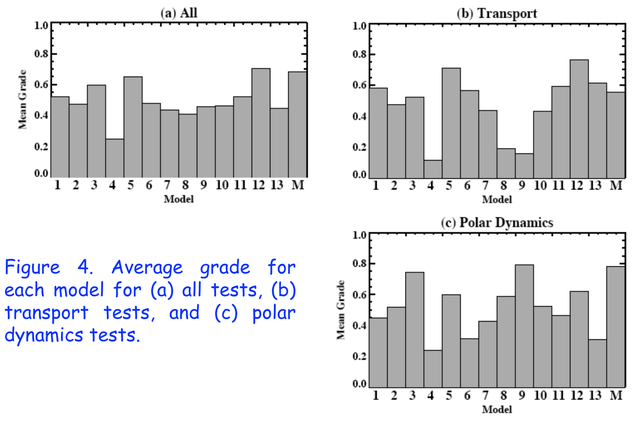

Figure 4. Average grade for each model for (a) all tests, (b) transport tests, and (c) polar dynamics tests.

Sensitivity to Observations

Several choices need to be made in the above analysis. One is the choice of observations. The sensitivity to source of observations is examined for the temperature diagnostics, where data is available from different meteorological centers. Fig. 5 shows (b) Temp-NP is not sensitive to analyses used, (b) Temp-SP has larger sensitivity but ranking of models unchanged, and (c) Temp-Trop is very sensitive (however, UKMO analyses have know 1-2 K warm bias).

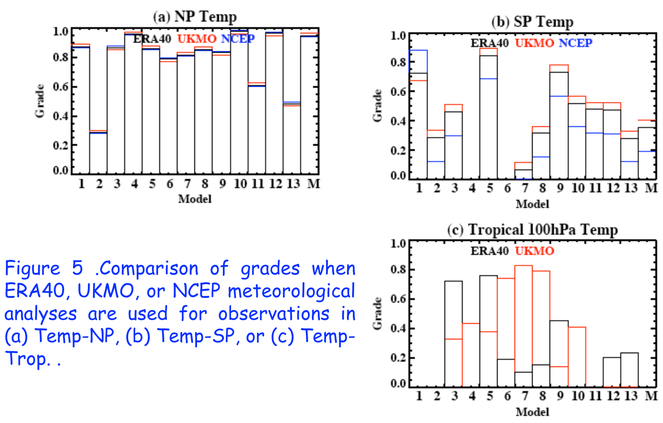

Figure 5 .Comparison of grades when ERA40, UKMO, or NCEP meteorological analyses are used for observations in (a) Temp-NP, (b) Temp-SP, or (c) Temp-Trop. .

Sensitivity to Grading Metric

Another choice is the grading metric used. An alternative metric is the t-statistic. However this can not be applied to all diagnostic as some lack long enough data records for calculation of observed variance.

For idealized case where model and observations have the same data length and same variance it can be shown that

$$t=\sqrt{\frac{n}{2}}n_g(1-g)$$

This can used to estimate t and statistical significance from the metric g. For example, model is statistically different from observations at p% level if g<g^* where

$$g*=1-\sqrt{\frac{2}{n}}\frac{t_p}{n_g}$$

and t_p is critical value for two-sided t test.

Although above relationship is exact only in idealized case, Fig. 6 shows it is a good approximation, at least for some of the tests considered here. Results presented are not likely sensitive to choice of g as the metric.

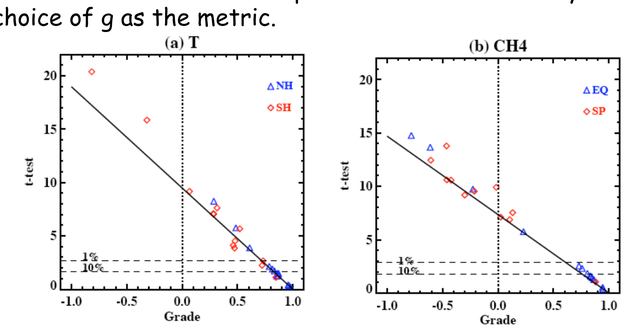

Figure 6. Comparison of t-statistic and g for (a) Temp-NH, Temp-SH. And (b) CH_4-EQ. CH_4-SP. The solid lines shows the theoretical relationship shown above.

Weighting Ozone Projections

The assignment of grades enables relative weights to be assigned to ozone projections from different models. A single performance index can be formed for each model:

$$\bar{g}_k=\frac{1}{W}\sum_{j=1}^{N}w_j g_{jk} \qquad W=\sum_{j=1}^{N}w_j$$

These indices can then be used to form weighted multi-model mean and variance of ozone predictions:

$$\hat{\mu}_X=\frac{1}{\sum\bar{g}_k}\sum_{k=1}^{M}\bar{g}_k X_k. \qquad \hat{\sigma}^2_X=\frac{\sum\bar{g}_k}{(\sum\bar{g}_k)^2-\sum\bar{g}_k^2}\sum_{k=1}^{M}\bar{g}_k(X_k-\hat{\mu}_X)^2.$$

Critical choice is the weights used to form the single model index. We have tried a variety of weights, and for the diagnostics and ozone projections considered there is, generally, only small differences between weighted and unweighted multi-model mean projections, see, e.g., Fig. 7.

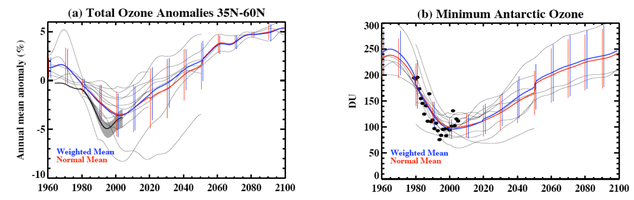

Figure 7. Temporal variation of (a) annual-mean total ozone anomalies for 35-60N and (b) minimum Antarctic total ozone for individual models (black), unweighted (red) and weighted (blue) mean of all models. Model performance indices are based on the average transport grade (Fig 4b).

Concluding Remarks

Analysis quantifies several features noted in Eyring et al. (2006).

Provides benchmark for evaluation of CCMVal-2 models.

Generally only small differences between weighted and unweighted mean ozone projections.

Issues:

Diagnostics of other key processes (e.g. chemistry), and variability as well as mean.

Source and uncertainty of observations.

Optimum set of diagnostics, and relative importance of each diagnostic.

More information: See Waugh & Eyring (2008) or E-mail Darryn Waugh waugh@jhu.edu

Acknowledgements

We thank the CCM groups for providing their model data to the CCMVal Archive and the CCMVal community for helpful discussions of this concept at the CCMVal 2007 workshop Co-ordination of this study was supported by the SPARC CCMVal project.. We thank the British Atmospheric Data Center for assistance with the CCMVal Archive. This work was supported by grants from NASA and NSF.

References

Douglass, A.R., et al., Choosing meteorological input for the global modeling initiative assessment of high-speed aircraft, *J. Geophys. Res., 104*, 27,545-27,564, 1999.

Eyring, V., N., et al.: Assessment of temperature, trace species, and ozone in chemistry-climate model simulations of the recent past, *J. Geophys. Res., 111*, D22308, doi:10.1029/2006JD007327, 2006.

Waugh D.W., & V. Eyring, Quantitative performance metrics for stratospheric-resolving chemistry-climate models, *Atmos. Chem. Phys. Discuss.*, 8, 10873-10911, 2008.