Atmos. Chem. Phys., 9, 9101–9110, 2009
www.atmos-chem-phys.net/9/9101/2009/

Atmospheric Chemistry and Physics

Comment on "Quantitative performance metrics for stratospheric-resolving chemistry-climate models" by Waugh and Eyring (2008)

V. Grewe and R. Sausen

Deutsches Zentrum für Luft- und Raumfahrt, Institut für Physik der Atmosphäre, Oberpfaffenhofen,
82230 Wessling, Germany

Received: 15 May 2009 – Published in Atmos. Chem. Phys. Discuss.: 26 June 2009
Revised: 23 November 2009 – Accepted: 25 November 2009 – Published: 1 December 2009

Abstract. This comment focuses on the statistical limitations of a model grading, as applied by D. Waugh and V. Eyring (2008) (WE08). The grade g is calculated for a specific diagnostic, which basically relates the difference of means of model and observational data to the standard deviation in the observational dataset. We performed Monte Carlo simulations, which show that this method has the potential to lead to large 95%-confidence intervals for the grade. Moreover, the difference between two model grades often has to be very large to become statistically significant. Since the confidence intervals were not considered in detail for all diagnostics, the grading in WE08 cannot be interpreted, without further analysis. The results of the statistical tests performed in WE08 agree with our findings. However, most of those tests are based on special cases, which implicitely assume that observations are available without any errors and that the interannual variability of the observational data and the model data are equal. Without these assumptions, the 95%-confidence intervals become even larger. Hence, the case, where we assumed perfect observations (ignored errors), provides a good estimate for an upper boundary of the threshold, below that a grade becomes statistically significant. Examples have shown that the 95%-confidence interval may even span the whole grading interval [0, 1]. Without considering confidence intervals, the grades presented in WE08 do not allow to decide whether a model result significantly deviates from reality. Neither in WE08 nor in our comment it is pointed out, which of the grades presented in WE08 inhibits such kind of significant deviation. However, our analysis of the grading method demonstrates the unacceptably high potential for these grades to be insignificant. This implies that the grades given by WE08 can not be interpreted by the reader. We further show that the inclusion of confidence intervals into the grading approach is necessary, since otherwise even a perfect model may get a low grade.

1 Introduction

Waugh and Eyring (2008) (WE08) applied a set of performance metrics to climate-chemistry models (CCMs) aiming at quantifying their ability to reproduce key processes relevant for stratospheric ozone. These performance metrics are used to calculate a quantitative measure of performance, i.e. a grade. These grades are employed to illustrate the ability of individual models to simulate individual processes and to identify general deficiencies in modelling key processes. These grades are further applied to weight individual CCM projections of the ozone layer to derive a weighted multi-model mean projection.

There is no doubt that the general approach, i.e. the model validation and grading, provides an important contribution to scientific questions regarding stratospheric ozone. However, the approach relies on the way the grading is performed and hence requires a statistical sound definition of the grading. Although the authors have discussed some statistical considerations, these considerations do not have any implications on the grading, e.g. no confidence interval for the grading value is given. Moreover, effects like uncertainties in the observational data are not included in most of the gradings: the grading formula (details see below) includes a parameter σ_{obs}, which is defined for some diganostics as the interannual variability and for others as an uncertainty

Correspondence to: V. Grewe
(volker.grewe@dlr.de)

due to the measurement. However, this does not replace the need for consideration of confidence intervals. Our results suggest that even the qualitative results, obtained in WE08 will change drastically when these effects are included in the grading.

In the next section, we are addressing the questions related to errors in model and observational data: "What statistical implications do these errors have on a model grading?" and "What statistical implications do they have on the difference of two model grades?". In Sect. 3, we define the general statistical terms, which we use. In Sect. 4, we present examples, which are aimed to clarify the shortcomings of the grading. In Sect. 5 the implications for a grading are discussed, when statistical significance levels are included in the grading approach. This illustrates the difference between the information on model performances presented in WE08 and the statistically sound information.

2 What is a grading?

Generally, a grading means "How well does a test object represent a certain reference value?". It consists of two parts, a test of the object against a reference value and a relation between the outcome of the test and a grade. That is exactly what the first two sentences of the abstract of Waugh and Eyring (WE08 in the following) is about: "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. The same metrics are used to assign a quantitative measure of performance ("grade") to each model-observations ..."

So there are two basic questions: "When does a test object represent the reference value?" and "How to derive a grade from a difference between the test value and the reference value?".

It is important to separate these two questions. It is necessary to find a method dealing with either question. This is one of the main reasons, why the methodology applied in WE08 does not provide the information it was designed for.

2.1 When is a model result representing an observation?

In this case the test objects are results from climate-chemistry models and the reference value is a certain observation.

The reference value itself is not precisely known, basically for four reasons:

1. uncertainties in measurement techniques,
2. uncertainties in methodology,
3. representativity for a certain region and time, and
4. representativity for a climatological value.

The first point summarizes all uncertainties associated with the measurement techniques, e.g., the precision of a measurement. The second one is more related to the processing of the measured data to derive the physical quantity, e.g., retrieval algorithms. The third one describes an uncertainty, which is related to the temporal and spatial coverage of the measurements. The sampling of data by satellite measurements might be restricted to clear sky conditions, a certain local time or a latitude-longitude-time relation. If vertical profiles or certain height information are used, these might only represent a certain height region, weighted with a kernel function, or the vertical localisation is given within an uncertainty range, only. These three types of uncertainties describe an uncertainty which is related to an observation for a certain area and time period. For simplicity reasons let us assume that the observations are given with a mean value μ_{obs} and an uncertainty range expressed by a standard deviation $\sigma_{\text{obs}}^{\text{unc}}$. Note that a bias also may occur, which complicates the whole picture.

In the case of a quality assessment of climate-chemistry models a further uncertainty has to be regarded in addition, namely climate variability. Because of the interannual variability, a climatological mean value can only be determined within a confidence interval. Here we assume that this variability can be expressed by a standard deviation $\sigma_{\text{obs}}^{\text{iav}}$. We further assume, unless stated otherwise, that the values are Gaussian distributed, which is in many cases not correct, e.g. for Northern Hemisphere temperatures. Note that the application of t-statistics also implicitely assumes a Gaussian distribution.

To summarize, all of these uncertainties limit the accuracy to which a climatological value from any observation can be determined. The uncertainties and errors are discussed in WE08 and the grading approach includes a variability measure, which is based on either the interannual variability or a measurement uncertainty. In Fig. 5 of WE08 the uncertainty of the grade with respect to the used observational data set (ECMWF versus UKMO) has been impressively demonstrated. However, these findings are not systematically included in the grading, i.e., from a grade 0.2 it cannot be determined, whether the grade is low because the model is incorrect or the observational data are badly estimated.

Figure 1 gives an illustration for the comparison of two 10 year model data sets (blue and green) and a 10 year observational data set (red). All are produced with computer generated random numbers for Gaussian distribution (black line) with expectation 0 and standard deviation 1. Note that this example neglects any uncertainties in the observational data, i.e. $\sigma_{\text{obs}}^{\text{unc}}$=0 and just takes into account an interannual variability.

In this example the underlying probability distribution is identical for the "model" and "observational" data. Hence model and the reality are identical, implying that the model is representing the reality perfectly. However, the 10 realisations for either "model run" and the "observations" differ.

The conclusion from this example is that from both, the observational data and the model data, the underlying probability distributions have to be estimated and compared in a statistical manner. This implies that at first a decision has to be made on the accuracy of the statements. I.e. what is the error that is tolerated in the decisions or in the uncertainty of the grade. Then the estimates for the probability distributions have to be compared and a decision can be made whether the model represents the reality or not. Note that in the example the assumed distribution is Gaussian, the estimates for the expectation (which is 0 in the example) is the sample mean value (differing from zero) and for the standard deviation the sample standard deviation of the data.

2.2 When is a model better than another?

Model grades are used and will be used to rank models. Grades condense a complex context into a single number. However, as shown in Fig. 1, every intercomparison can only lead to a grade within a certain error range, which depends on a large number of parameters. Hence two models (Fig. 1) might get two very different grades, however with errorbars that are so large that the grades themselves do not differ statistically. Therefore a grade itself is meaningless, unless an estimate for the uncertainty is given. Two model results are given (blue and green) in the example above (Fig. 1). They are realisations (random samples) of the same random variable (X), which in this case has a normal distribution with expectation $E(X)$=0 and standard deviation $S(X)$=1 (N(0, 1)). However, their gradings differ: Model X has a grade of 0.77 and Model Y of 0.46, when applying Eq. (4) of WE08:

$$g = \begin{cases} 1 - \frac{1}{n_g} \frac{|\mu_{\text{model}} - \mu_{\text{obs}}|}{\sigma_{\text{obs}}}, \text{ if} & \frac{1}{n_g} \frac{|\mu_{\text{model}} - \mu_{\text{obs}}|}{\sigma_{\text{obs}}} \leq 1 \\ 0, & \text{else} \end{cases} \qquad (1)$$

where μ_{model} and μ_{obs} are the sample mean values of the model sample and observational sample, respectively. σ_{obs} is the sample standard deviation of the observational data and n_g a factor (here: 3 as in WE08), which relates the difference of sample mean values of the observations and model data to n_g times the sample standard deviation of the observational data. This metric has been applied in earlier studies (Douglass et al., 1999; Kawa et al., 1999).

Following WE08, Model Y would be better than Model X in this example, i.e. this example clearly shows the limitations of the grading methodology as applied in WE08. In the following, terms and definitions are given. These form the basis for a systematic analysis, which is performed to show that the example, given above, is not an extreme outlier, but a representative example. The methodology can also be used as a basis for analysing further grading approaches.

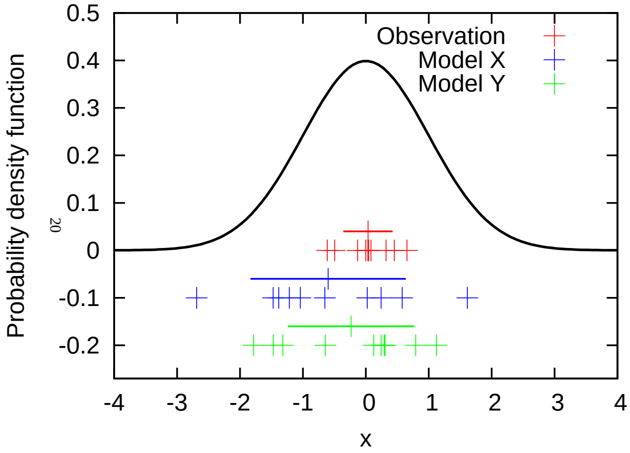

Fig. 1. Three random experiments (red, blue, and green) with 10 realisations each. A random number generator with with a Gaussian distribution and an expectation of 0 and standard deviation 1 (black line) is applied to generate these numbers. I.e. this represents reality ($E^r = 0$ and $S^r = 1$). The calculated sample mean values ($\mu_{\text{modX}} = 0.0345$; $\mu_{\text{modY}} = -0.5998$; $\mu_{\text{obs}} = -0.2339$) and sample standard deviations ($\sigma_{\text{modX}} = 0.3902$; $\sigma_{\text{modY}} = 1.234$; $\sigma_{\text{obs}} = 1.005$) for the three realisations are shown on top of each row. The red one is taken as an observational dataset, the other two as results from 2 model experiments. Hence the mean values of the observation, Model X and Model Y will all converge to 0 (=expectation of the normal distribution) with increasing sample size.

3 Terms and definitions

In the following, X, Y, Z denote random variables representing a given diagnostic for two models "Model X" and "Model Y" and observations. We are considering N realisations of either Model X and Model Y, i.e. we have 2 samples, $X_1,...,X_N$, and $Y_1,...,Y_N$, and a sample $Z_1,...,Z_M$ for the observations with sample size M.

We assume that the random variables have normal distributions, with expectations $E(X)$, $E(Y)$ and $E(Z)$ and standard deviation $S(X)$, $S(Y)$ and $S(Z)$. To be consistent with WE08, we denote the sample means of $X_1,...,X_N$, $Y_1,...,Y_N$, and $Z_1,...,Z_M$ as μ_{modX}, μ_{modY}, and μ_{obs}. Note that in this case μ is *not* the expectation. In analogy, the sample standard deviations are given by σ_{modX}, σ_{modY}, and σ_{obs}. Further, we denote E^r and S^r the real expectation and real standard deviation, describing the real atmosphere. Hence a model is perfect if $E(X)$=E^r and $S(X)$=S^r and observations are perfect if $E(Z)$=E^r and $S(Z)$=S^r.

Further GX and GY denote random variables of the grading of Model X and Model Y, and $GX_1,...,GX_K$ and $GY_1,...,GY_K$ are samples of the grades of Model X and Y, respectively. The samples have a sample size K each and are calculated on the basis of samples of the random variables X, Y and Z with sample sizes N for the models and M for the observations. The sample mean values are given

Table 1. Overview on the four statistical tests performed to determine thresholds for grades. Details are described in Sect. 3. X and Y are random variables representing 2 models and, Z is the random variable representing observations, and GX and GY are the respective random variables for the model grades. $E(\bullet)$ and $S(\bullet)$ denote the expectation and standard deviation. E^r and S^r denote expectation and standard deviation of the reality. The threshold are calculated by inverting the given equation and the underlying pdf is calculated with Monte-Carlo simulations with the given conditions.

			Information on the determination of the threshold	
	H_0 Hypothesis	Threshold	Equations	Conditions
Model differs from observations	$E(GX)=E(GZ)$	$g^{\text{obs}}(p)$	$P(GX \leq g^{\text{obs}}(p))=p$	$E(X)=E(Z)$ $S(X)=S(Z)$
Model differs from reality	$E(GX)=E(GZ)$	$g^{\text{real}}(p)$	$P(GX \leq g^{\text{real}}(p))=p$	$E(X)=E^r$ $S(X)=S^r$
2 Models differ (perfect obs.)	$E(GX)=E(GY)$	$\Delta g(p)$	$P(\lvert GX-GY\rvert > \Delta g(p))=p$	$E(X)=E(Y)$ $S(X)=S(Y)$ $E(Z)=E^r$ $S(Z)=S^r$
2 Models differ (imperfect obs.)	$E(GX)=E(GY)$	$\Delta g(p)$	$P(\lvert GX-GY\rvert > \Delta g(p))=p$	$E(X)=E(Y)$ $S(X)=S(Y)$ $E(Z)=E^r+\alpha\times S^r$ $S(Z)=\beta\times S^r$

by μ_{GX} and μ_{GY}, i.e. μ_{GX} and μ_{GY} are the mean values of K grades of either Model X and Y with N samples of the random variables X and Y (models) and M samples of the random variable Z (observational data). Note that for one specific model run (as in WE08) the sample size K equals 1. The expectations of the random variable GX and GY are $E(GX)$ and $E(GY)$.

In the following, we introduce the statistical tests, performed within this study. An overview can be found in Table 1. A model is then statistically different from the observations, if the null hypothesis H_0: “Model and observations have the same expectation” can be rejected and the alternative hypothesis H_1: “Model and observations are different” can be accepted. For a model that does not differ statistically from observational data, i.e. for $E(X)=E(Z)$ and $S(X)=S(Z)$, we determine the threshold value $g^{\text{obs}}(p)$ for which the probability that $GX \leq g^{\text{obs}}(p)$ is p, i.e. $P(GX \leq g^{\text{obs}}(p))=p$, where p is the probability that H_0 is erroneously rejected. (We use p=1% and 5% in the following.) As an example, we consider one realisation GX_1 of the random variable GX, e.g. as in WE08. We reject the null hypothesis and regard the model as statistically significantly different from the observations, if the grading GX_1 of Model X is smaller than $g^{\text{obs}}(p)$.

A model is then considered imperfect, if the null hypothesis H_0: “Model and reality have the same expectations” is rejected and the alternative hypothesis “Model and reality have different expectations” is accepted. Hence we determine the threshold value $g^{\text{real}}(p)$, where $P(GX < g^{\text{real}}(p))=p$.

Note that these two tests seem to be very similar, however they have different implications. They differ in the expectation and standard variation of the random variable Z, which are $E(Z)$ and $S(Z)$ in the first case and E^r and S^r in the second case. In general, observational data are erroneous, which means that $E(Z) \neq E^r$ or $S(Z) \neq S^r$. This has an impact on the grading, which we will describe in detail below.

Two model gradings are then statistically different, if the null hypothesis H_0: “Grade of Model X and grade of Model Y are equal” can be rejected and the alternative hypothesis H_1: “Grades of Model X and Model Y are different” can be accepted. Hence we determine the threshold value $\Delta g(p)$, where $P(|GX-GY| > \Delta g(p))=p$, with $E(X)=E(Y)$ and $S(X)=S(Y)$.

The statistical tests given in WE08 (their Sect. 2.2) are special cases identical to those described above, with the assumptions $E(Z)=E^r$ and $S(Z)=S^r$ and $\sigma_{\text{modX}}=\sigma_{\text{modY}}=\sigma_{\text{obs}}$. Our analysis will show that there is no disagreement between our findings and the results presented in WE08 with respect to the special cases. However, in general, the confidence intervals for the gradings are much larger than for these special cases. An inclusion of confidence levels into the grading will change the grading results drastically.

4 Examples

4.1 A perfect model and perfect observations

Let us assume that we have a perfect model and that we know perfectly the regarded diagnostic of the reality. Further we assume that the values of the individual years (here: $N=M=10$ years) have normal distributions. Since model and observations are considered to be perfect in this case, the model's and observational's expectations and standard deviations are equal: $E(X)=E(Z)=E^r=0$ and $S(X)=S(Z)=S^r=1$.

The expectation of the grading ($E(GX)$) is then:

$$E(GX)=1-\frac{1}{3}\frac{|E(X-Z)|}{S(Z)} \neq 1-\frac{1}{3}\frac{|E(X)-E(Z)|}{S(Z)}=1,$$

in general.

We estimate $E(GX)$ by means of a Monte Carlo simulation. We perform K=100 000 realisations of GX. Each has random samples of X and Z with sample size N (=10). The resulting probability density function is given in Fig. 2. The derived sample mean value for the grading is μ_{GX}=0.87

(since K is large μ_{GX} already converged to $E(GX)$) and the median is 0.88. They hence differ remarkably from 1. Calculating the p=5% (1%) percentile from the frequency distribution (Fig. 2) gives a value of $g^{\text{obs}}(p)=g^{\text{real}}(p)$=0.65 (0.5) for this case.

For illustration purpose we additionally assume an uniform distribution of the random variables X and Z, with values between -1 and +1. The resulting pdf for GX is the blue line in Fig. 2. It only slightly differs from the Gaussian distribution.

4.2 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, 2008).

Let us assume that (as above) the reality can be described by an expectation E^r and a standard deviation S^r. If we had perfect observations the sample mean value μ_{obs} would be close to E^r, for a large number of observational data (M). Let us now assume that the observational data have an error. We express this error by an offset in the expectation and standard deviation: $E(Z)=E^r+\alpha\times S^r$ and $S(Z)=\beta\times S^r$. I.e. the observations have an error expressed by a multiple (or fraction) of the standard deviation, which is the interannual variability in the case of annual mean data.

An uncertainty in the expectation (=α) of 50% to a factor of 3 of the standard deviation is a reasonable assumption, as we will show by 2 examples in the following. For midlatitude (35° N–60° N) total ozone columns, the interannual variability is in the range of 5% and the differences between the various datasets (Ground-based, SBUV, NIWA, GOME) are around 2–3%, which is around 50% of the interannual standard deviation (see WMO (2006) p. 3.11) . Lary and Aulov (2008) presented distributions of HCl measurements, e.g. for January at 450 K to 590 K isentropic levels and between 49 and 61° N. Differences between 3 measurement systems are around 0.3 ppbv, whereas the interannual variability for HALOE January values is in the order of 0.1 ppbv, which gives a factor of α=3.

Figure 3 shows the mean values, 5% and 1% percentiles of the grading parameter. The coordinate (0, 1) represents the perfect observation, i.e., the example in Sect. 4.1. Clearly, the grading of the perfect model depends on the quality of the observational data. An increasing error in the expectation and hence the sample mean value leads to a reduction of the grading value. If the standard deviation in the observational data is lower than in reality, the grading value for the perfect model is also reduced. Whereas the model gets a better grading, if the standard deviation of the observation is larger than in reality. The 5% and 1% percentiles (Fig. 3, mid and bottom) are decreasing to grading values of lower than 0.2, if either parameter has a 50% uncertainty. Hence, the case without errors in the observations or whenever error estimates are not available, gives a good estimate for an upper boundary of the threshold (=left side of the confidence interval) at which a grade becomes significantly different from that of a perfect model.

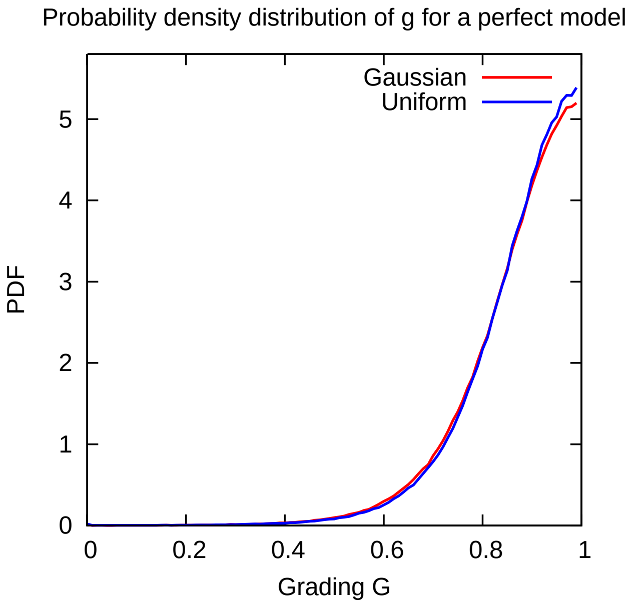

Fig. 2. Probability density function for the random variable GX (=grading) of a perfect model on the basis of perfect observations for a Gaussian (red) and a uniform (blue) distribution (between -1 and +1).

To summarize, allowing a 1% error margin means that all models with a grading of more than 0.1 have to be regarded to be perfect, for most of the observational data qualities regarded in this example.

4.3 Two identical models

Here the difference of two model gradings is investigated, i.e. we answer the question “Is Model X statistically different from Model Y?” (see also Sect. 3).

Let us first assume that we have perfect observations and two identical, but imperfect models, with expectation $E(X)=E(Y)=E^r+\alpha_{\text{mod}}\times S^r$ and standard deviation $S(X)=S(Y)=\beta_{\text{mod}}\times S^r$. (Both models are perfect for α_{mod}=0 and β_{mod}=1.) The expectation of either model grading is identical $E(GX)=E(GY)$ and the difference of both is 0.

In the example in Sect. 4.2, the expectation and standard deviation of the observations were overlaid with an error. Here, the same error approach is applied to the 2 models. The parameters α_{mod} and β_{mod}, are randomly chosen, but equal for the models. They cover a deviation of maximum 2 times S^r. The parameter range is smaller than in the previous example and actually describes small values for current

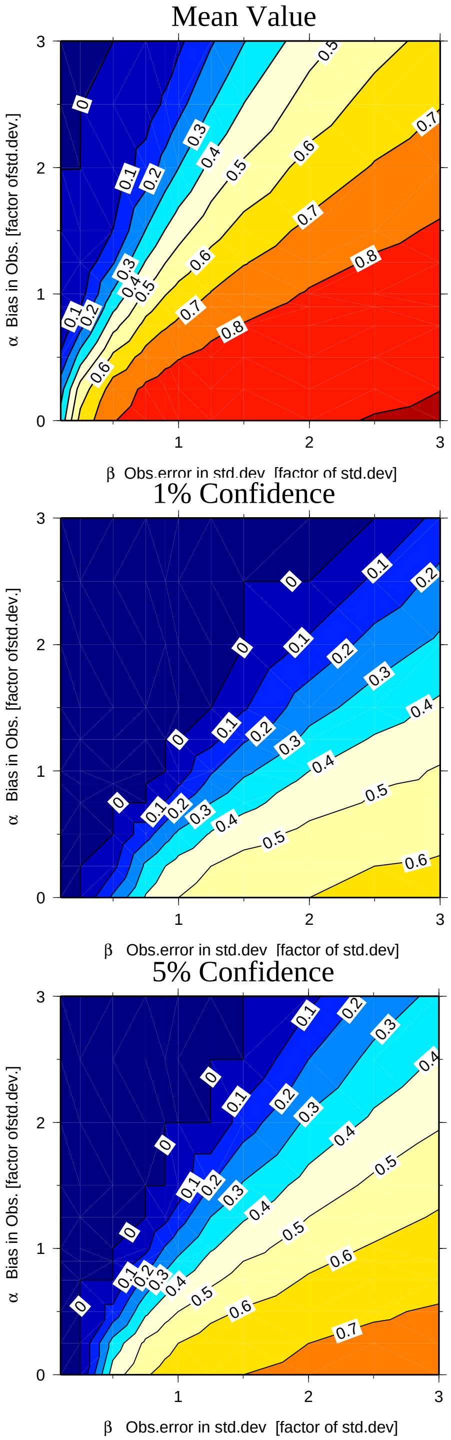

Fig. 3. Top: mean grading, i.e. expectation of the random variable GX (=$E(GX)$) (top), 5%- (mid), and 1%-percentiles (bottom) for a perfect model and imperfect observations. The error in the observations is defined by an offset in the expectation (y-axes) and a multiple of the standard deviation (x-axis). The offset is a multiple of the standard deviation. α=0 and β=1 represents perfect observations.

CCMs. For each of the 2 parameters 23 parameter settings were chosen in the given range. For each setting 10 000 iterations (=K) were calculated to estimate the probability density function of the difference in the two model grades GX and GY, which add up to more than 5 million iterations. As an illustration, Figure 1 shows one such an iteration, with N=M=10 values for the observations, and Model X and Y, each, though for perfect models. The frequency distribution of the 1% and 5% percentiles for the absolute difference in the two model grades are shown in Fig. 4. The mean 5% and 1% percentiles of the absoute difference for all regarded parameter settings are 0.33 and 0.42 (vertical lines). However, in 5% of the parameter settings 1% (5%) of the model differences are larger than 0.65 (0.51). And in 1% of the parameter settings 1% (5%) of the differences are larger than 0.72 (0.55), defining the 1%- and 5%-percentiles.

In Fig. 4 (bottom) results are presented with inclusion of imperfect observational data. The error is analogously considered: $E(Z)$=E^r+$\alpha_{\text{obs}}\times S^r$ and $S(Z)$=$\beta_{\text{obs}}\times S^r$. $\alpha_{\text{obs}}, \sigma_{\text{obs}}, \alpha_{\text{mod}}, \beta_{\text{mod}}$ are independently chosen in the range [0.5, 2]. The results are similar to those with perfect observation, except that the confidence intervals are significantly increasing. The distributions have longer tails.

This leads to the conclusion that based on their gradings, two models are not distinguishable with these assumptions, unless the difference is larger than 0.71 and 0.86 for perfect and imperfect observational data, respectively.

To summarize, these examples demonstrate that the statistical tests performed in WE08 are in agreement with our findings, however for special cases only. They give a threshold g^*=0.7 for N=11, $E(Z)$=E^r, $S(Z)$=S^r, and σ_{obs}=σ_{modX}, which roughly corresponds to our value $g^{\text{real}}(p$=0.05)=0.65 for N=10, $E(Z)$=E^r, $S(Z)$=S^r, however $\sigma_{\text{obs}}\neq\sigma_{\text{modX}}$.

Further they give a threshold value of 0.3 for a statistically significant difference in two model gradings at a 5% significance level, with the assumptions $N=11$, $E(Z)$=E^r, $S(Z)$=S^r, and σ_{obs}=σ_{modX}=σ_{modY}. Our corresponding value is $\Delta g^{\text{real}}(p$=0.05)=0.33 for N=10, $E(Z)$=E^r, $S(Z)$=S^r, however $\sigma_{\text{obs}}\neq\sigma_{\text{modX}}$=$\sigma_{\text{modY}}$.

The examples further show that the values for the special cases analysed in WE08 are misleading and that an adequate inclusion of observational errors change these thresholds. Without any further analysis of the uncertainties in the observational data, it cannot be decided whether a 95% confidence interval spans 1/3 or the whole grading interval [0, 1].

5 Consequences for the grading

In the last section we have investigated the reliability of the grading according to WE08. In this section we show its implications on the overall grading picture, i.e. on their Figs. 2 and 4. Applying the same procedure as in the previous sections, we randomly defined 16 diagnostics and 13 models.

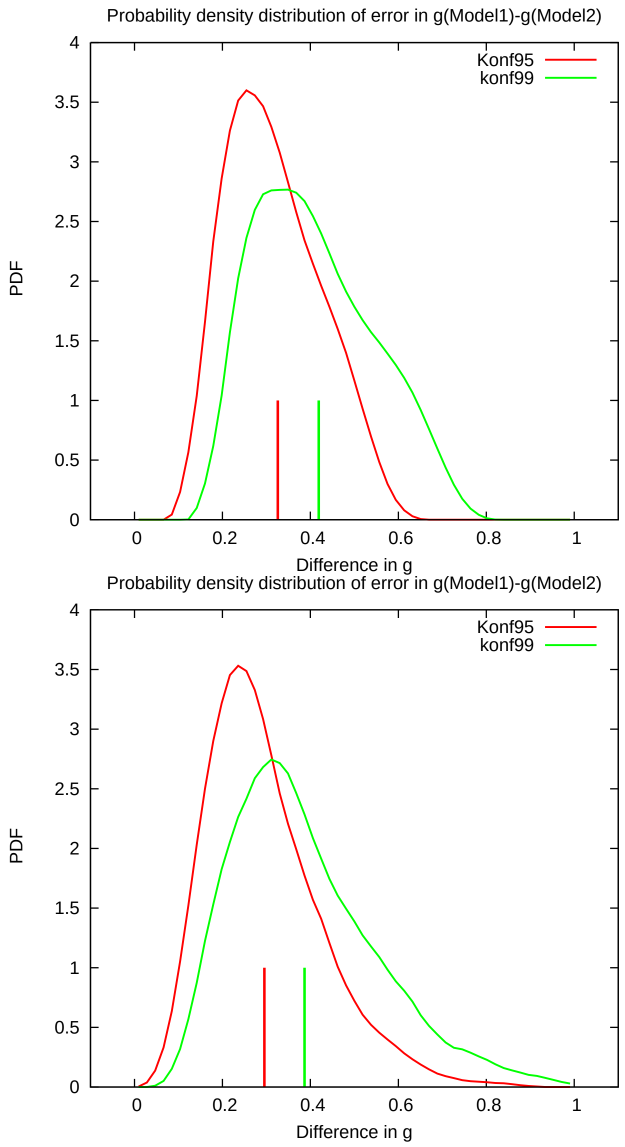

Fig. 4. Probability density distribution of the the 1% (green) and 5% (red) percentile for the absolute difference of two random variables $|GX-GY|$, i.e. the difference in the grading of two imperfect but identical models for perfect (top) and imperfect (bottom) observations. The vertical lines show the expectation of the 1% and 5% percentiles.

We then compare two grading approaches: the first is identical to that in WE08 and the second maps this grading to a 0 to 1 scale, where 1 is defined by the 5%-percentile of the grade from WE08. Any major qualitative differences occurring between these gradings imply that the grading of WE08 is not sound. The diagnostics are based on M=N=5 to 40 years of data, the expectation and standard deviation of the observations and models vary by randomly chosen factors α, β between 0.25 and 2 and 3, respectively. Hence the models have potentially a larger error than the observations. A detailed description of the parameters is given in the supplement material.

Figure 5 (left) shows the grading matrix in analogy to WE08, but for our random models and observations. The diagnostic 1 and 4 has for all models high grades, whereas the diagnostic 13 and 16 leads to low grades for all models. For all of the diagnostics, we have calculated the 5% percentile for the expected grade of a perfect model and the given imperfect observations. Figure 5 (right) shows in black all grades, which do not differ significantly from reality. For all other grades the distance of the grade to the confidence interval is taken as a deviation from the grade 1. Therefore, Fig. 5 (right) shows the significant model grades (s-grades), where a s-grade 1 indicates a model, which is not distinguishable from reality for the respective diagnostic. And all models with a lower s-grade do differ significantly.

This changes the picture of the grading considerably. For diagnostic 1, which is characterised by high grades, the s-grades are high, but half of the models are imperfect. Whereas for diagnostic 2 many models have low grades, but since the confidence interval is large, all models are not distinguishable from reality and hence get a perfect s-grade of 1.

The quality of model 1 is similar to the other models with respect to the grade (Fig. 5, left). However, taking into account the confidence intervals for the grades, this model becomes perfect with s-grades of 1 for all diagnostics. The qualitative difference in the grading g and statistical sound s-grade for our random models and observations implies that the grades given in Fig. 2 in WE08 are not reliable.

Figure 6 shows for the diagnostic 7 the grades of all models, comparable to Fig. 4 in WE08. We chose this diagnostic, because it is one of those showing a variability among the models in the grades and s-grades. We pick out model 1 and look for a significant difference to other models. Six models (red) do not differ significantly from reality and from model 1. Although model 5 and 6 (grey) are significantly different from reality, they do not differ significantly from model 1. Only the models 3, 7, 9, 11 (green) differ significantly from model 1 and they also differ significantly from reality. Hence a ranking of the models, which is suggested for a weighting of the multi-model mean is a quite difficult task, e.g. although the s-grades differ for model 1 and 5, both models do not differ statistically significant from reality.

6 Conclusions

In the paper “Quantitative performance metrics for stratospheric-resolving chemistry-climate models” by Waugh and Eyring (2008) a method was introduced, which converts the outcome of a diagnostic, i.e. a comparison of climate-chemistry model data and observational data, into a grade. A grading was applied to a number of diagnostics, leading to an overall model grade, which was proposed to be used as a weighting for a multi model mean.

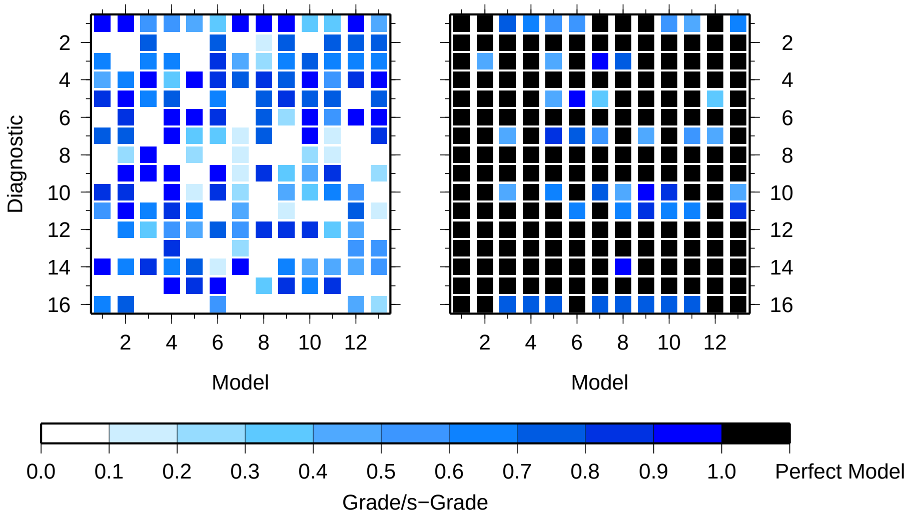

Fig. 5. Grading matrices for 13 random models and 16 random diagnostics for 2 different statistical interpretations of the grading Eq. (1). Model performances and observational data are randomly defined, but equal for both grading interpretations. Left: Grading applied without any further statistical considerations (as in WE08). Right: Grading with inclusion of 95% confidence intervals (s-grade). I.e. a model s-grade equals 1, if the model grade is larger than the threshold ($g^{\text{real}}(p)$, Table 1) for which a model that cannot be statistically significant distinguished from reality. (See text for details).

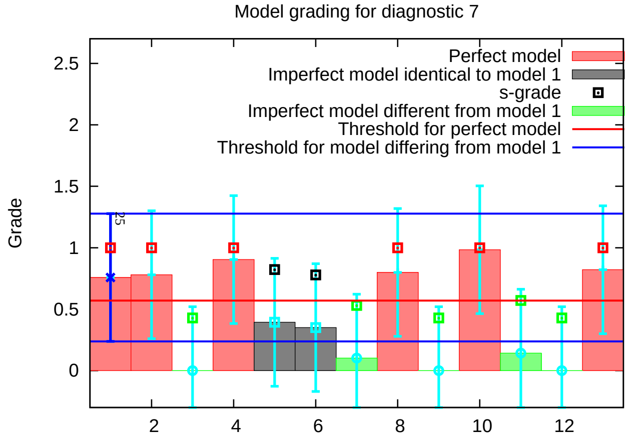

Fig. 6. Grading of the models for diagnostic 7. Horizontal lines mark the 5% percentile for a perfect model (red) and for significant difference of either model 2–13 to model 1 (blue). Models, which do not differ significantly from reality at a 95% confidence level are marked in red. Models, which significantly differ from model 1 are marked in green. Models, which do not differ significantly from model 1, but which differ significantly from reality are marked in grey. Errorbars indicate the 95% confidence interval for the grading difference to any other model, i.e. within these errorbars two models are equal. The squares mark the s-grades.

In this comment, we focus on the statistical basis for the grading, with two aspects, the statistical confidence in the grade itself, and the possibility to statistically distinguish two models with this grading. A summary is given in Table 2. Even if perfect observations could be performed, and a perfect model is applied, an expected grading of 0.87 is obtained for a ten year dataset, the 99%-confidence interval for the model’s grade is $[0.5,1]$. If we were not able to perform perfect observations, i.e. the observations have a bias in the order of σ_{obs}, the interannual variability, then this confidence interval is even enlarged to almost the whole range of the grade g $[0.01,1]$.

Two models differ statistically, if their grades differ by more than 0.33 and 0.42 for an assumed error of 5% and 1%. However, these are mean values for a range of possible imperfect observational data. In 1% of the regarded errors in observational data, a difference in the grade of more than 0.86 is needed to significantly distinguish two models. And note that no answer is yet given on how much the models differ. If a statistical significant minimum difference (e.g. 0.1, 0.2, 0.3, ...) is regarded for the difference of two models, then this requires confidence intervals for each chosen minimum difference. Hence a ranking of the models is hardly possible, and applicability for a multi-model mean is very limited.

In Fig. 2 in WE08 the grades of a number of models and diagnostics are presented. The grading does not include any uncertainty in the observational dataset for most performance metrics. The parameter σ_{obs} in their formula describes either an interannual variability or an uncertainty due to measurements. In the first case the measurements are implicitely regarded as perfect. And hence those are comparable to the example in Sect. 4.1. This implies that all models with a grade larger than 0.5 have to be regarded to be perfect. Lower model grades indicate a significant difference to the observational data. This is a qualitative statement and a further quantification on a statistically sound basis cannot be given.

Table 2. Overview on the results from the Monte Carlo simulations. Imperfect observations are defined by an offset in the expectation and a multiple in the standard deviation (Details see text).

	Percentile	
	5%	1%
Model differs significantly from reality (perfect observations)	$g < 0.65$	$g < 0.50$
Model differs significantly from reality (imperfect observations)	$g < 0.20$	$g < 0.01$
Two models differ significantly (perfect observations)	$\overline{\lvert\Delta g\rvert} > 0.33$	$\overline{\lvert\Delta g\rvert} > 0.42$
Two models differ significantly in 5% of perfect observations	$\lvert\Delta g\rvert > 0.51$	$\lvert\Delta g\rvert > 0.65$
Two models differ significantly in 1% of perfect observations	$\lvert\Delta g\rvert > 0.55$	$\lvert\Delta g\rvert > 0.71$
Two models differ significantly in 5% of imperfect observations	$\lvert\Delta g\rvert > 0.52$	$\lvert\Delta g\rvert > 0.69$
Two models differ significantly in 1% of imperfect observations	$\lvert\Delta g\rvert > 0.65$	$\lvert\Delta g\rvert > 0.86$

However, the observational data have uncertainties, which should be accounted for. A thorough re-analysis of the grading would imply an estimate of the observational errors and interannual variabilities of all used observational datasets, which has not been performed. If only a 25% or 50% uncertainty with respect to the standard deviation is taken into account for the mean value and the standard deviation, then the results for the random models presented here suggest that without any further consideration of the measurement uncertainties, we cannot decide whether most of the models presented in Fig. 2 in WE08 differ on a statistical basis. There might indeed be cases where differences in grades are large enough to assume that those differences are statistically significant. But from the analysis performed so far we simply do not know. It has to be noted that observational error estimates are not available for all of the addressed parameters. Often uncertainties from retrievals or uncertainties associated with the representativity for a certain region and time are rarely addressed. An educated guess might be necessary. Omission of such uncertainties because they are not known would implicitely imply that these uncertainties are zero, which is probably more unrealistic than an educated guess.

The statistical tests, which we performed are in agreement with those performed in WE08. Their tests are however only special cases, which assume perfect observations and that the interannual variability in the observational data equals the interannual variability in the model data. The generalisation of the statistical test with the inclusion of observational errors and differing interannual variability in the observational and model data clearly shows a distinct difference in the results, e.g. grading confidence levels. Moreover the inclusion of the statistical findings into the grading approach was not performed in WE08, which limits the interpretation of the results, since it is not clear, which gradings are statistical significant or which model gradings differ statistically from each other. I.e., the results presented in Fig. 2 in WE08, e.g., for the diagnostic Temp-Trop do not reflect the findings presented in Fig. 5c in WE08, showing the large dependency on the observational dataset.

We have not addressed the comparability of the grades for different diagnostics so far. If an observation has a very large error, i.e. little valuable information can be drawn from this observation the grade after WE08 would be large, if σ_{obs} refers to this uncertainty. On the other hand, if there is a very accurate observation, already a rather small offset between the model and the observation may result in a low grade. A generalisation, e.g. which model bias in terms of multiple of the interannual variability is tolerable, might not work. One may decide in certain cases independently from the statistics, which deviation from the observations should be tolerated.

A further challenge, which has not been addressed so far, is the robustness of a multi-diagnostic grade. In this comment, the grading properties were investigated on the basis of one diagnostic, only. If more than one diagnostic is taken into account, the variability of the individual grades has to be combined somehow with the confidence intervals to provide an overall model grade with an uncertainty range.

The evaluation of models is an important part of model development. Multi-model approaches are the only way to address questions, which are of high importance to politics and society. Model grading helps to better understand model differences and determine specific model shortcomings. Hence a statistical sound grading is absolutely necessary. We propose a detailed verification of any further grading methodology, e.g., on the basis of Monte Carlo simulations. And we further strongly suggest not to consider a grading approach in the way it was done for any further multi-modelling study. In detail, we propose for any future grading (a) to either calculate, estimate, or rely on expert judgement for all of the errors 1–3 described in Sect. 2.1, as well as for the interannual variability; (b) to include these uncertainties in the grading approach such that if the model data cannot be statistically distinguished from reality then and only then the grade is 1; (c) to also include these uncertainties in the determination of grades lower than 1 (e.g. $1-x$), such that for a given significance level, model data and reality differ significantly by at least a certain value, which corresponds to some value x in the grading.

Acknowledgement. We like to thank Darryn Waugh for a first discussion of this comment, which was very helpful for us in trying to understand the different views on a "grading".

Edited by: P. Spichtinger

References

Douglass, A. R., Prather, M. J., Hall, T. M., Strahan, S. E., Rasch, P. J., Sparling, L. C., Coy, L., and Rodriguez, J. M.: Choosing meteorological input for the global modeling initiative assessment of high-speed aircraft, J. Geophys. Res., 104, 27545–27564, 1999.

Kawa, S. R., Anderson, J. G., Baughcum, S. L., Brock, C. A., Brune, W. H., Cohen, R. C., Kinnison, D. E., Newman, P. A., Rodriguez, J. M., Stolarski, R. S., Waugh, D., and Wofsy, S. C.: Assessment of the Effects of High-Speed Aircraft in the Stratosphere: 1998, NASA-Report, NASA/TP-1999-209237, 1999.

Lary, D. J. and Aulov, O.: Space-based measurements of HCl: Intercomparison and historical contex, J. Geophys. Res., 113, D15S04, doi:10.1029/2007JD008715, 2008.

Waugh, D. W. and Eyring, V.: Quantitative performance metrics for stratospheric-resolving chemistry-climate models, Atmos. Chem. Phys., 8, 5699–5713, 2008, http://www.atmos-chem-phys.net/8/5699/2008/.

World Meteorological Organisation (WMO): Scientfic Assessment of Ozone Depletion, Geneva, 2006.