**This reply is written on behalf of all authors.**
**For a better understanding of the reply we cited the comments in black and give our answers in red.**

**Generally, we appreciate any effort, which leads to a better evaluation process of CCMs. The comment clearly indicates, where agreements and where different scientific views exist. We hope that our response clarifies the raised issues.**

**Two things have to be mentioned:**

1. **Waugh and Eyring are right that our main conclusion, namely**
   ***that the calculated grades are insignificant,***
   **has to be rephrased, namely**
   ***that neither RS09 nor WE08 have shown, which grades can be interpreted in a way that the model significantly deviates from reality on a certain confidence level. Our general analysis shows of the grading method shows that there is a high potential for the grades to be insignificant. This implies that the grades given by WE08 can not be interpreted by the reader.***
2. **Other arguments are discussed in detail in the response, but we do not see the necessity to change our methodology or specific results. However, in most cases there is the need to clarify some issues in the manuscript.**

---

Citing from the comment:

The comment by Grewe and Sausen (hereinafter "GS09") examines the statistical significance of the grading calculations performed in our 2008 paper on "Quantitative performance metrics for stratospheric-resolving chemistry-climate models" (Waugh and Eyring, 2008; hereinafter WE08). We are pleased to see that the authors are continuing the conversation on model grading. We agree with them that more analysis is needed on the statistical robustness of the grading metric used in WE08, and any other grading metric, and also that there needs to be more consideration of confidence levels and uncertainty in observations (e.g. by considering observations from multiple instruments and platforms) in model grading. Furthermore, we think the framework they use is a useful way to examine these issues. However, we object to their conclusions that the method presented in WE08 is not leading to statistical significant grades and that differences between two models is hardly significant. These negative statements about WE08 are not justified or supported by their analysis. We think the unaccounted uncertainty in the observations that are used in GS09 are unrealistically large, and cannot be used to refute our analysis. In addition, there are many cases where the grades in WE08 Figure 2 for two models are statistically different even if based on the GS09 estimates of significance.

Reply:

We are pleased to see a common understanding on the necessity of developing a statistically robust grading. We are also pleased to see that the methodology and mathematics applied is not doubted.
Darryn Waugh and Veronika Eyring give two reasons why they believe that our conclusions are incorrect, although the methodology is correct:
a) The regarded uncertainties are not representing the reality and are chosen unrealistically large.
b) There are model grades presented in WE08, which differ statistically significant.

Yes, we agree that we have not shown in our comment what the uncertainties are for each individual diagnostic and grade and we have not shown, whether the grades presented in WE08 differ significantly or not.
We have not done this and WE08 have not done this! That is exactly our point of the comment.
We requested an analysis, which shows what information is included in each individual grade. Otherwise we compare numbers (grades), for which we do not know what they actually mean.
And yes, this is a lot of work, but it is necessary to be performed, if we want to have reliable grades. Without giving confidence levels, numbers (grades) don't tell anything. It is not a two step approach, first calculating grades and secondly (i.e. in an upcoming paper), refining the methodology and including confidence intervals. It cannot be separated, since the grades are not interpretable without the second step.

We are aware of the problem that the uncertainties are difficult to estimate. A conservative estimate in the sense of 'what is the minimum confidence interval' that we can achieve is when we assume to have perfect observations. Even then, our (revised) conclusions hold. And therefore the choice of the uncertainty ranges (their argument a) is not limiting the (revised) conclusions.

However, Darryn Waugh and Veronika Eyring are right in their argument that we have not shown that the grades, which they presented, differ significantly or not. Nor they did! So the conclusion is that we simply do not know unless we do the analysis. But the general analysis we have performed shows that the chosen grading approach has a very limited content of information. We will rephrase our conclusions in that way (details see below).

In our analysis, we have adopted a two step approach. First, we analyzed the uncertainties in the case of perfect observations and then estimated a range of uncertainties, which we regarded to reflect a realistic range. And yes, we agree that without a thorough analysis of the observational uncertainties and the implications on the confidence level, we cannot conclude, whether the grades are meaningful or not. This is exactly the point. We do not know, neither from WE08 nor from GS09, whether the grades have 95%-confidence intervals in the order of ±0.1, ±0.5, or ±1.0. Our argument is that a thorough analysis would have been required in WE08 to obtain this information. In our paper we gave an estimate of these quantities for perfect observations and a range of imperfect observations. We further analyzed the potential impact on a grading matrix as presented in WE08 and showed that there is no systematic agreement between the grades as defined in WE08 and grades, which include a confidence interval (s-grades). Therefore our conclusion is that the uncertainties of the grading in WE08 are not sufficiently investigated. As a consequence the reader of WE08 has no chance to decide whether the different grades are meaningful or not.
We might have overstated the conclusions. We agree that there are gradings, which reflect the model performance correctly and b) is probably right. But how can the reader of WE08 decide, which of the gradings and which difference between gradings are meaningful or not? This is exactly our point: In the revised version, we rephrased these statements as indicated further below:

Citing from the comment:
1. Statistical significant grading: In numerous places the authors claim that the WE08 analysis does not lead to statistical significant grading, that difference in grades between models is not significant, and that the methodology does not provide the information it was designed for (e.g., in the abstract and in statements on pg 14142, line 6, 7 and 11; pg 14144, line 5; pg 14152, line 17; pg 14155, line 18, and pg

14156, line 12). These allegations against our paper need to be thoroughly justified. However, GS09 do not perform any analysis of the actual diagnostics / grading performed in WE08. They consider synthetic random datasets, and assume large uncertainties in the observations rather than considering the uncertainties in the actual observations used in WE08. For some of the grades in WE08 there may be low statistical significance because of large unaccounted uncertainties, but they have not shown this. More importantly, in order to justify their generalized negative statements against our paper, GS09 would need to show that this is the case for all grades presented in Fig. 2 of WE08. They don't do this, and there are numerous cases where there are significant differences between grades.

<u>Reply:</u>

A couple of things should be made clear at this point:

a) WE08 used a statistical method (grading formula) for 16 diagnostics and 13 models, without analyzing the confidence interval for each of the grades.

b) WE08 did not include uncertainties in the measurement data for <u>all</u> grades. (Note that interannual variability is not an uncertainty of the measurement!).

We analyzed the confidence interval for observations, which are assumed to have no error and found that the confidence interval is large. This finding is valid for all grades without any consideration of the specific observation, except for the number of degrees of freedom.

Our conclusion is that the confidence intervals are too large, as that one could regard the method to be statistically robust. Since for most diagnostics the confidence interval is not calculated by WE08, we question the reliability of the grades in general.

We think that it is the opposite WE08 should have shown that the grades in their Fig. 2 are reliable, i.e. give the, e.g. 95%-confidence interval. We believe that it is sufficient to question a methodology (here: grading approach) if it can be shown in general that this method has the potential to lead to drastically misleading results. – That is what we have done in GS09.

<u>We rephrased the sentences:</u>

14142, line 6, 7:

*"Monte Carlo simulations show that this method is not leading to statistical significant gradings. Moreover, the difference between two models is hardly significant."*

INTO

*„Monte Carlo simulations show that this method has the potential to lead to large 95%-confidence intervals for the grade. Moreover, the difference between two model grades has often 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.*

14142, line 11

*"Without these assumptions the grading becomes basically insignificant."*

INTO

*„Without these assumptions, the 95%-confidence intervals become even larger. Examples have shown that the 95%-confidence interval may even span the whole grading interval [0,1]. Without any further consideration of the confidence interval, the results in WE08 cannot be interpreted.Neither we nor WE08 have shown, which of the grades presented in WE08 can be interpreted in a way that the model significantly deviates from reality on a certain confidence level. Our general analysis the grading method shows that there is a high potential for the grades to be insignificant. This implies that the grades given by WE08 can not be interpreted by the reader.*

14152, line 17

*„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 drastically. “*

INTO

"*„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]. “*

"

14155, line 18,

*„… suggests that basically none of the models presented in Fig. 2 in WE08 differ on a statistical basis“*

INTO

*„… suggests 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. Note that some models might have grades, which are statistically significantly different, but from the analysis performed so far we simply do not know.“*

Citing from the comment:

One clear example is $Cl_y$, shown in Fig. 1 of WE08. Here the uncertainty used in the grading is the uncertainties in the observations, so the fact that the observations are not perfect has been accounted for. Furthermore, there are models that clearly have unrealistic values (e.g., models with peak $Cl_y$ around or less than 1 ppb in polar regions which are in fact more than 5$\sigma$ from the mean observations) and grades of zero. These model-observation differences are statistically significant and these models are also statistically different from other models that simulate peak $Cl_y$ around 3 ppb (see Review #2 for more discussion the significance of model-observation differences for $Cl_y$). Another example is the tape recorder, where again the uncertainties in the observations are used in the grading. This case is discussed in the comment by Strahan et al. Even in cases where the uncertainties in the observations are not included in the grading there are significant model-observation and model-model differences, see point 3 below.

Reply:

Yes, we also think that the extreme grades obtained with the Cly diagnostic are significant and we will re-phrase the conclusions as discussed above. However, the important point is that for this conclusion additional information has to be taken into account and the results from WE08 do not stand alone. The additional information is that we assume that the interannual variability of the Cly values is small. From Eyring et al. (2006) we get a minimum in the relative standard deviation of less than 1% (AMTRAC, Fig. 12b) to around 25% for SOCOL. This convinces us that the Cly concentration is far to low for the models E39/C and SOCOL. However without this additional information the grades alone are not interpretable, since it may well be that the simulated standard deviation is very large and the observed interannual standard deviation is very small. In this case the mean value might not be significantly different. The variability might than be totally wrong. However, the grades are a test of the mean value based on its concept. Therefore a low grade in the Cly diagnostic can only be interpreted correctly with the additional information on the modeled and observed standard deviation. And actually the low grades are very likely to be statistically significant, but we only know this taking into account additional information, which is not included in the grading function and only implicitly in Figure 1 in WE08.

The same holds for the all other diagnostics like the tape recorder.

Citing from the comment:

2. Observations: In their calculations GS09 assume there are large uncertainties in the observations used in the grading that are not accounted for in WE08. If there are uncertainties of factors of 2 and 3 times the interannual standard deviation then the significance of the WE08 grading will be low (if this is not accounted for). However, as discussed in the comment by Strahan et al. and the review of Reviewer #2, the uncertainty in most observational datasets used in process-oriented diagnostics is much smaller than this. GS09 give only two examples, one with 50% and another with factor 3, to justify their parameters. Moreover, neither of these observations was used in WE08, and to make statements about WE08 results it is necessary to consider the same observations and diagnostics.

Reply:

Generally, we have shown that even with a 0 uncertainty of the observations the grades have the potential to be insignificant. The investigation of uncertainties in the range of 50% to 200%-300% was chosen to give an idea on what might happen with the inclusion of measurement uncertainties.
Therefore, our main conclusions are basically unaffected by the choice of the uncertainty range.

It is not necessary to consider the exact same observations and diagnostics in WE08 to obtain an understanding of the general limitations of the grading formula. Our point is that if we can show that generally the grading leads to insignificant results for a lot of parameter settings and even for an observational uncertainty, which equals 0, then this clearly points at a shortcoming of the approach, questioning the approach in general. The reader simply does not know which of the models is able to simulate a mean value, which does not differ significantly from reality. And that is what we have shown in general. For specific diagnostics it has to be shown, how large the grading threshold for a perfect model is. Without this information the grades are meaningless.
But yes - - again, we have not shown this for any specific diagnostic used in WE08. Nor have WE08 shown the opposite.

Citing from the comment:

A quick consideration of the observations used in WE08 indicates that the unaccounted uncertainties are much smaller than assumed by GS09. First, consider the analyzed temperatures used for the polar dynamics tests in WE08 we can assess the uncertainty by comparing the different meteorological analyses, as GS09 did to assess the uncertainty in ozone and HCl. For Temp-NH the interannual standard deviation (sigma) for the 20 years of ERA40 analyses is 2.9 K, but the difference between means from UKMO and NCEP and ERA40 is only -0.14 K and 0.6 K, respectively. This corresponds to $\alpha$ less than 5%! For Temp-SH the corresponding values are sigma=3.13 K, UKMO-ERA40=-0.45 K, and NCEP-ERA40=1.5 K which corresponds to $\alpha$ = 14% and 48%. So $\alpha$ is larger in the SH than in the NH but still less than 50%. In fact the larger value results from the ERA40 to NCEP bias and, as stated in WE08, the NCEP data have a well-documented bias (of order 1-3 K) in the Antarctic lower stratosphere during winter-spring. So 48% is an overestimate of $\alpha$ for ERA40 in the SH, and using around 15% would actually be more appropriate. Another example where the uncertainties in observations are much smaller than GS09 claim is the U-SP diagnostic. This is based on quantities shown in Fig. 2 of Eyring et al. (2006): the difference between date of transition to easterlies at 20 hPa (where the grading is performed) from the three meteorological analyses shown in this figure is much smaller than the interannual standard deviation, and using the GS09 method this corresponds to $\alpha$ <20%.

Reply:
We assume that the difference between NCEP and ERA40 is 0.06 K not 0.6 K for Temp-NH?
These are examples for which $\alpha$ is definitely small. In this case our results for $\alpha$=0 might be more appproriate. However, $\alpha$ is definitely larger for TEMP-TROP as can be seen in Fig. 5c (Model 7) in WE08, where to different observations lead to grades of 0.1 and 0.85.

Citing from the comment:
Unaccounted uncertainties of the magnitude of the three examples above will only have a relatively small impact on the values of g required for statistical significance, and the estimates of confidence levels in WE08 would be an underestimate, but reasonable. There are however cases where the uncertainties are larger. One case is the tropical tropopause temperature diagnostic (Temp-Trop). Here the difference between the meteorological analyses is similar to the interannual variability (see Fig. 7a of Eyring et al., 2006). However, this difference is shown in WE08, and we highlight this as a case where the grading is sensitive to the observations used. In this case many of the differences in grades shown in WE08 will have low statistical significance, but there are still cases where the differences are significant (i.e., cases where g<0).

Reply:
How can you be sure that a grade less than 0 is statistically significant, if the test has actually not been performed?
Page 5076 in WE reads:
*„The wide range of grades for all diagnostic tests shows that there are no tests where all models perform well or all models perform poorly. However, the majority of models perform well in simulating north polar temperatures and NH and SH heat fluxes (mean grades over all models are larger than 0.7), and, to a lesser degree, mid-latitude age (mean grade greater than 0.6), see Fig. 3.“*

Hence the authors do interpret all grades without taking into account a statistical significance.

Citing from the comment:
3. Incorrect conclusions in GS09: Even aside from the suitability of the uncertainties in observations assumed by GS09, some of the statements are not justified. In Fig. 6 of WE08 that shows a comparison of the t-statistic with the grading metric g for the diagnostics Temp-NP and Temp-SP as well as for the transport diagnostics using methane ($CH_4$-EQ and $CH_4$-SP), there is a huge spread among the models in the metric g and also the t-statistic for these four diagnostics. It is clear that there are many models that are significantly different from the observations and pairs of models that are significantly different. Even using the values listed in Table 1 of GS09 (which are based on calculations assuming only 10 years and uncertainties as large as a factor of 2) there are statistically significant differences between grades shown in Fig. 6 of WE08. These are only four diagnostics, but these examples show that the broad generalized negative statements made in the abstract and elsewhere of GS09 about lack of significance of grades in WE08 are not correct.

Reply:

The diagnostics Temp-NP and Temp-SP are probably the most reliable, since uncertainties from the measurements and the interannual variability are small. Let us look in more detail into the diagnostic SP-Temp, i.e. Fig. 1 (bottom, left) of Eyring et al., (2006). The two orange curves (CMAM and UMSLIMCAT) have a negative temperature deviation with respect to ERA-40 of around 5K and get a grade of around 0.4 to 0.5. The interannual standard deviation is around ±3K for the ERA-40 data and the difference to UKMO around 2-3 K. The question is what would have happened if the UKMO data were taken into account and the interannual variability of the model is ±5 K instead of ±3K in the observations?

t-value for ERA-40 data: 5/SQRT((3^2+3^2)/20) = 7.45

t-value for UKMO and larger model variability: 3/SQRT((3^2+5^2)/20) = 2.3

This would increase the grade from 0.44 to around 0.67 and shift the model in Figure 6a from far beyond the 10% t-test line to the area between the1% and 10% line. And hence the mean value of the model and observational data won't be significantly different. (Note that the 1% and 10% lines are probably swapped in Figure 6a?)

Therefore, even for the more reliable cases (e.g. Temp-SP), the robustness of the grade is questionable.

Citing from the comment:

Also, we do not see the justification for the statement in the conclusions that "If only a 25% or 50% uncertainty with respect to the standard deviation, then the results for the random models presented here suggests that basically none of the models presented in Fig 2 of WE08 differ on a statistical basis". One reason it is hard to see the justification is that GS09 don't quote numbers for significant values of g for α = 25 to 50% percent (they focus on α between 50% and 2 to 3) and they don't say what they think is required for two models to differ (i.e., how many grades need to differ for two models to differ). However, if we again take the numbers in Table 1 of GS09 as representative numbers then |Δg| > 0.5 is required for two models to differ (using 5% for significance) for a given diagnostic. A quick inspection of Fig. 2 of WE08 shows that there is a very large number of cases where grades from two models for the same diagnostic differ by 0.5 or more. Hence, in contrast to what GS09 claim, there are many cases where the grades for two models are statistically different based on their own estimates of significance.

Reply:

In Section 5 we have also included lower values of alpha, like 0.25. Waugh and Eyring are right that this has not been sufficiently described. We have revised the text accordingly. And, as for the discussion above, we changed our conclusions: We didn't show explicitly, which model grades differ significantly. However, taking into account the large difference of at least 0.5 needed to get a significant difference: "We concluded that neither we nor WE08 show, which model grades differ significantly. However, the difference in the grades depends on the interannual variability and measurement uncertainties. The results indicate that only a few model grades may differ significantly."

Let's take as an example again the diagnostic TROP-TEMP: Figure 2 suggests that model 3 and 5 differ from all other models, perhaps with the exception of model 9, by taking a difference of 0.5 into account. However in Figure 5, these models do not show any significant difference (with a 0.5 threshold), when using the UKMO data.

Citing from the comment:

4. Misleading statements: In the Introduction of GS09 it is stated that "no confidence interval for the grading value is given in WE08" yet WE08 quote values of the grades required for two models to be different or for a model to be different from observations at 5 and 1% levels (e.g., g needs to differ by 0.3 for two models to differ at 5% level when the number of years (N) is 11). Furthermore, in Section 2.2 of GS09 it is stated that a random model with grade 0.77 would be better than a random model with grade 0.46 and that this clearly shows the limitations of the grading methodology as applied in WE08. However, as mentioned above WE08 quote a value 0.3 for difference between grades to be significant at 5% level for N=11. So the difference between models with grades 0.77 and 0.46 is right at this level for significant (and actually not significant if recalculated for N=10), and so is not really a clear evidence of the limitation of the WE08 methodology.

Reply:

Correct, the sentence is misleading and is now rephrased. As said at other places in GS09, WE08 did a statistical test, however, only for special cases. And this information is not reflected in Figure 2.
The general expectation is that a model, which gets less than half of the possible points can't actually be good. This is supported by sentences like that quoted above (green text). But actually this is only because of the short sampling period. The example given for the models with grades 0.77 and 0.46 will get grades equal to 1 for longer simulation periods.

Citing from the comment:

5. Constructive analysis: We think that rather than performing and presenting results for random selection of N and $\alpha$ from a range of values it would better to calculate and show the critical values listed in Table 1 of GS09 for specified values of N and $\alpha$. For example, it would be useful to plot how the minimum $|\Delta g|$ for two models to differ at 5% level increases as $\alpha$ increases, for several different values of N. Such calculations could be used to examine the significance of grades in WE08 (once further estimates of uncertainty in observations used by WE08 are obtained), and could also be used by other scientists applying the same metric to different diagnostics / models.

Reply:

Actually this can't be done, since the statistics depend on the interannual variability of the model data, too. The method proposed here (s-grades) gives a solution to the problem. The grading formula may still not be perfect.

Citing from the comment:

Although we have issues with many of the statements in GS09 we think the framework they present is useful and should be pursued. If they repeat their analysis for the actual diagnostics used in WE08 together with best estimates of the uncertainties in the various observational datasets then they will be able to make justified comments about the significance of the WE08 grades.

Reply:

We are grateful for the open discussion by Darryn Waugh and Veronika Eyring. We would like to add two things:

- The results obtained with assumptions of having perfect observation (first part in GS09) can be seen as an indication for the lower boundary for the confidence intervals. And the analysis gives an overview on the information this grading approach is able to provide. But, and therewith Waugh and Eyring are right that we can not explicitly draw the conclusion that any individual grading given in WE08 differs significantly from a grading a perfect model would get. However, the results indicate that the grading procedure in the way it is applied in WE08 is not statistically robust.
- Actually we are planning as a next step to revise the calculation with 'real data'. However this takes a while.

Citing from the comment:

The limitations of the grading used in WE08 are summarized, and the need for further research stated, in the conclusions of our paper. We therefore again strongly support research on this topic and will be happy to see more statistically robust metrics developed and applied to future CCM studies.

Reply:

Yes, limitations are summarized in WE08. However, we do not think that research groups are aware of the abilities of this grading and the grading will be used for many future studies without further development of the grading formula (we are not talking about the diagnostics and including more diagnostics, but to revise the grading itself). We still believe that the use of a grading, without an inclusion of confidence intervals in the grading itself is insufficient and we still recommend not to use it, without the inclusion of statistical confidence levels.

*References:*

Eyring, V., et al. (2006), 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.