Reply to reviewer #2 on behalf of all authors:

The referee basically agrees with the methodology and the outcome, but disagrees with the conclusions, because we have not explicitly tested the identical diagnostics as in WE08.
The referee is totally right and we have changed the text accordingly.
Beyond that, we are working on a grading methodology, which is beyond the scope of this comment, but rather topic of an independent paper.

In the following, please find a summary of the changes we made concerning this point.
(This can also be found in the reply to Waugh and Eyring's comment)

We will rephrase the sentences:
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 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.

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.“

Minor points:

1. right. Changed.
2. ok
3. The number was already given in line 5: 10 years, Additionally we now repeat it in line 12 ("N=10")
4. Correct, this was meant as an illustration, revised.
5. changed
6. I think misleading is correct, since in some cases it even might be pessimistic.