Reply to reviewer #1 on behalf of all authors.

Thanks for the constructive comments.

1. We include a discussion of the problems with uncertainties in the discussion section, as suggested.
2. We agree with the referee. However, the grades are used for CCM evaluation. A further discussion on the use for CTMs may distract the reader from the main statistical problems with this grading approach. Therefore, we prefer not to start this discussion at this point.
3. That is actually a good point, which is not raised so far. Our proposal is to include in the sigma, the interannual variability only and to calculate the 95%-confidence intervals for each pair diagnostic/model separately. We add a short discussion to the conclusions. Actually, the accuracy meant in the text refers to the tolerance of the error in the grading, which of course is related to the tolerance in the error of the observational data. This question would need further considerations, which are beyond the scope of our paper.

Minor:

1. changed accordingly
2. correct! deleted.
3. correct, and this is probably the way forward. However, we do not see the need to discuss this further, in order to avoid distraction from the main points, since it does not help with the principle statistical problems the grading has.
4. Great comment! We didn't dare to start this discussion. But a hint is now given in the text.
5. The idea is that Model X and the random variable *X* are linked. The same holds for Model Y and the random variable *Y* and observations and random variable *Z*.
6. rephrased. Should be clear now.
7. Right. We agree that there are many other potentially interesting questions, however we refrain from including this in the text in order to avoid that the main message becomes hidden.
8. Rephrased
9. Good point! No, actually, the number of iteration range from 5 to 40, as stated in the text.

The indicated typos were removed.