

Answers to Referee #1

At some places, the text could be a bit more concise and the punctuation could be improved.

We will look through the entire text and improve this aspect whenever possible.

Minor comments:

Abstract: The first third of the abstracts reads more like an introduction. Could be reformulated more concisely.

We agree. The first three sentences can be deleted. The fourth sentence would then need a slight modification.

Ln 2 (and at other places). Not so clear what "alternative" means. Maybe better something like "forced climate models under equally valid assumptions for boundary conditions"

We will clarify the meaning of "alternative" simulations whenever relevant. The referee's suggestion is fine, but "alternative" could also be thought of as alternative choices of parametrization of small-scale processes in one and the same climate model under exactly the same boundary conditions. Different parametrizations could make the same basic model having different climate sensitivities, and thus it could be interesting to evaluate if one parametrization leads to a better fit to observational (proxy) data than another.

Page 2633, ln 21ff (and §5). It would be interesting to include also individual members of the ensemble. Sure, the ensemble mean will filter out better the forced signal, but maybe the method can also be used to define the actual "best" simulation (e.g. for data assimilation purposes).

The U_R and U_T statistics can certainly be calculated for individual ensemble members, and the calculated values will naturally not be identical. Actually, Fig. 3 and Fig. 5 in Part 2 of our series (Hind et al., Clim. Past, 8, 1355-1365, 2012) compare results for individual members and the whole ensembles using pseudoproxy data. However, if one member has "better" statistics than another then this is just a random result, which cannot be explained by anything else than that the chosen initial conditions by chance happen to cause a better fit between simulated and observed data sequences. We have not considered the possibility to use our metrics for selecting "the best" individual member within an ensemble where members only differ by their initial conditions, and we would not recommend to do so. On the other hand, if some user would like to do this in a data assimilation study, then this may be fine. After all, the calculation of our metrics is an objective procedure based on a deliberately formulated statistical model and some reasonable assumptions. However, there is no implication that a particular ensemble member with a lower D^2 will be any better in prediction of the future climate. It could be an interesting study to investigate in a pseudoproxy setting how our metrics could be of use in a data assimilation context. This, however, goes beyond our current plans.

Ln 2642, §4.2: a relatively long paragraph coming to a very simplistic conclusion could be summarized in two sentences or so.

Agree. We can simplify and shorten this section.

Page 2653, ln 1: maybe worth mentioning that Servonnat also point to relatively long lead times the climate system needs to show response to forcing. This could be realized differently in models and the real climate.

Yes, the time lag between forcing and response could be differ among different models and also compared to reality. This is actually one of the reasons why our framework is of practical interest – to evaluate if one model fits the observed climate better than another. We can refer to this point made by Servonnat at some relevant place when we revise our text.

Appendix B, page 2660, ln 21ff: “In the situation that a particular forcing effect is present . . . both with and without forcing. . .”: this sounds a bit strange, how can there be a forcing effect without forcing?

Thanks for pointing out this. We did not write clearly enough what we meant. Of course, it cannot be a forcing effect if there is no forcing. The situation we consider here is one where we compare models which share one particular forcing (or a set of several forcings) but they differ by means of another (additional) forcing. We will make this clear in our revised text.

Stockholm, November 2, 2014

Anders Moberg, Rolf Sundberg, Håkan Grudd, Alistair Hind