Point-to-point replies to Reviewer comments for manuscript esd-2017-70

Comment by the authors:

Please find below point-to-point replies to the comments of the anonymous Referee. Responses are given in red. Changes in the revised manuscript according to the comments are also given in red.

General comments: -----------------

1. I suggest changing the order of results shown in section 4.1 (Fig. 2 & 3). It seems more logical to me - and hence easier to follow for the reader - to present the skill analyses starting with a reference forecast that would be expected to be rather easy to outperform (climatological forecast) before using reference forecasts that are a priori known to contain skill (the uninitialzed simulations). The results in section 4.2 then continue this logical order by using even more sophisticated reference forecasts, the initialized global predictions.

Answer: We have changed the order of the results in section 4.1 as suggested by the Reviewer. Hence, Fig. 2 now shows the MSESS using the climatology as reference, while in Fig. 3 the uninitialized historicals are used. We have changed the text in section 4.1 accordingly in the revised manuscript.

2. I have the feeling that you did not fully get the point of my comment 3 of last round's review, that is the issue of ensemble size and its impact on skill. I agree that this issue has to be considered seperately from the issue of few initializations (rather use this term than "sample size", from my understanding the "sample size" comprises both, the number of initializations times the ensemble size). Let's focus on the ensemble size issue here. I still insist on my point of view that this problem of an ensemble-size dependent bias is known (and solutions or let's say workarounds). This is a mathematical issue of the various skill metrics as they are and has nothing to do with the origin of the actual data (synthetic, weather or climate prediction, global or regional). In the end the bias itself and its behaviour with a growing number of ensemble members depends only on the signal-to-noise-ratio (which crucially depends on the variance of the targeted variable). Anyway, as I wrote in my previous review: I appreciate your effort of tackling this issue (many studies do not, and probably aren't even aware of this problem). However, your results do not add anything new to the scientific literature regarding the existence of this bias. That is why, I oppose against your research question 3 (Does [...] skill depend on ensemble size?). To resolve this issue, I suggest the following: Please reformulate your research question 3 to something like "How does ensemble size impact regional decadal prediction skill?" It's a small change but this "how" makes a difference from my point of view. And please, carefully revise your text in a sense that it does not sound anymore as if the basic question (dependence yes or no) is something that has never been addressed before. Maybe you can write something like "the ensemble-size dependent skill bias has never been demonstarted based on regional decadal climate predictions before".

Answer: Following the Reviewer’s suggestion we have reformulated research question 3 in the revised manuscript. Further, we have revised the text in section 4.3 in the sense suggested by the Reviewer.

3. Regarding your research question 4: As shortly mentioned above, please refrain from naming this issue as a matter of sample size. The sample size in the end is determined by the ensemble size and the number of initializations. Please make clear that you address the issue of few initializations here!

Answer: We have changed this research question as suggested. Further, we have revised section 4.4 and parts of the discussion accordingly.

Specific comments: ------------------

1. Page 2, line 15: I suggest including a reference to Eade et al. (2012) related to the prediction of extremes.

Answer: We have included the reference in the revised version.

2. Page 2, lines 31-33: Please split in two sentences. First one to end after "...techniques". And I suggest to replace "outstanding" with "exceptionally".

Answer: We have reformulated this part as suggested by the Reviewer.

3. Page 3, line 7: "aspread" should be "spread".

Answer: It is now changed to “spread”.

4. Page 3, line 12: Replace "for the decadal predictability" by "regarding skill". The reference to the decadal timespan comes later in the very same sentence. And "predictability" actually is to be differentiated from "prediction skill" but this is another discussion and unfortunately done inaccurate by many colleagues.

Answer: We have reformulated it in the revised manuscript.

5. Page 3, lines 21-22: Please reformulate reserach questions 3 and 4 according to the suggestions made in my general comments.

Answer: We have reformulated research question 3 and 4 as suggested by the Reviewer.

6. Page 3, line 32: As already mentioned by another reviewer during the discussion stage of this manuscript (and answered correctly by you), ocean temperature and salinity are NOT taken from NCEP/NOAA reanalysis. Please pay attention to this issue and describe the initialization procedure correctly.

Answer: We clarified the procedure in the revised manuscript: “The first generation (baseline0; Müller et al., 2012, Matei et al. 2012) is initialised with oceanic conditions from an experiment, where surface fluxes from the NCEP/NOAA reanalysis (Kalnay et al., 1996) were assimilated into the ocean model MPI-OM. The anomalies of ocean temperature and salinity from this experiment were then used to initialize the decadal hindcasts in the coupled model."

7. Page 4, lines 7-9: I suggest removing the sentence regarding the downscaling from this paragraph. It's almost exactly repeated in the following paragraph (lines 15-17) which is specifically dedicated to the downscaling.

Answer: We agree with the Reviewer, that this is a repetition. However, we prefer not to remove the whole sentence, as it contains abbreviations which are used throughout the manuscript. Instead we have reformulated this sentence such that the downscaling is not mentioned anymore.

8. Just as a matter of curiosity: Did you test whether it makes a difference for your skill estimates if you take (interpolated) winds from ERA-reanalyses directly instead of using the ERA-driven CCLM-simulation as a reference?

Answer: This is an interesting point. However, we did not test this in our study. We can only speculate, but we assume that it may indeed impact the skill estimates when using winds directly from ERA-reanalysis, as some physical mechanisms are not captured when "simply" interpolating ERA-Interim winds to a finer grid.

9. Page 4, lines 27-28: It is not correct that "historical" simulations are forced ONLY by aerosol andgreenhouse gas concentrations. There is a number of other external forcings that are prescribed. Please rephrase accordingly.

Answer: That is of course correct. We rephrased the sentence in the revised manuscript: "With this aim, a 10-member ensemble of uninitialised MPI-ESM-LR historical runs started from a pre-industrial control simulation are used, which use observed natural and anthropogenic forcings (e.g. aerosol and greenhouse gas concentrations among others) for the period 1850-2005 (e.g. Müller et al., 2012)."

10. Page 4, lines 31-32: What is the interpolation method you used? Did you test for alternative interpolation methods and the related impact on skill? It's quite a range of resolutions you are using here. Given that you interpolate from much coarser but also from (slightly) higher resolution (featuring a rotated grid) to the 0.25deg-grid, I guess bi-linear interpolation for all datasets would be the most appropriate solution. In any case, please indicate your the method chosen by you.

Answer: We have actually used the bi-linear interpolation method in our study to interpolate all datasets to the E-OBS grid (0.25°x0.25° resolution). We have clarified this in the revised version. With respect to the second question, we have tested different interpolation methods in other studies, and found that the impact on the skill estimates are negligible. TODO: Please check my reformulation and comment in the manuscript.

11. Page 5, lines 1-2: How were the anomalies calculated for the hindcasts, that is to mean, how did you calculate a climatology from the hindcasts? This question may sound stupid, but the devil is in the detail and it might even be, that your hindcasts feature drifts (even though it is anomaly initialization). I think the latest recommendation in this respect is to define a baseline period that is covered by the same number of initializations for all lead times and then calculate a climatology for every single lead time separately. This is not possible for you given that you have initializations only five-yearly. I don't ask for a change in your approach here (whatever you did) but please describe precisely how you did it.

Answer: We did exactly what we have written in the manuscript. To calculate the anomalies, we removed the mean over the period 1961-2010 from the hindcasts and the observations, respectively. Tests using a different reference periods did not reveal changes in the results presented here.

12. Page 5, line 4: I suggest replacing "mainly" by "partly". You do spatial averaging over the regions only for the results presented in tables 1-3. However, the number of plots where you did not perform spatial averaging is much higher.

Answer: We have changed it to “partly” as suggested by the Reviewer.

13.Page 5, line 21: Remove the reference to Goddard et al (2013) here. They were not the ones defining the MSESS.

Answer: We have replaced “Goddard et al.” by “Murphy, 1988” in the revised version.

14. Page 6, line 1: Two issues regarding the MSESS-formula:

14.1. Please replace the vertical bars (indicating the calculation of an absolute value) by brackets. Essentially it doesn't make a difference here but, you should follow the derivation presented by Murphy (1988).

Answer: We have replaced the vertical bars by brackets.

14.2. More generally; i wonder if it is useful to present a formula for the MSESS that holds only for the climatology as reference forecast. You present results, too, that are based on other reference forecasts, so it may be better to stay with the basic definition of MSE and MSESS as provided on page 5 already. If you want to present some decomposition, I suggest to stay with the more general one, I wrote down in my previous review. And given that these derivations are provided by other papers already, I don't see the need to indicate some derivation of the conditional bias in your paper from my point of view, you could directly provide the formula for the conditional bias as it is (see below). But please spend a few sentences, describing explicitely that the MSESS (partly) depends on the correlation but also on the condictional bias. This would be an important step, helping the reader to establish links between your MSESS- and ACC-results. This is what I asked for in my previous review.

Answer: We followed this suggestions an added the more general form of the MSESS decomposition in chapter 3 and extended the description of the dependency of the MSESS on the correlation and the conditional bias.

15. Page 6, line 4: Now the vertical bars are definitely wrong. Your tables contain negative values for the CB, too. So, it should be brackets (or nothing).

Answer: We have replaced the vertical bars by brackets as suggested.

16. Page 6, line 14: The correlation (ACC) is not only independent from the mean bias but also from the variance of the specific target variables.

Answer: We thank the Reviewer for this hint. We have added this information in the revised version.

17. Page 7, line 1-3: Please avoid using "bias" in this context. This might be misleading. Maybe just replace by "deviation", or something similar.

Answer: We have replaced it by “deviations” as suggested.

18. Page 7, line 19: Delete "of the hindcasts and decadal predictions".

Answer: We have deleted it in the revised manuscript.

19. Page 7, line 20: Replace "more reliable" by something like "better" or similar. Reliability has a specific meaning in forecast verification that is not meant here.

Answer: We agree that “reliable” is misleading here. We have now replaced it by “better predictions”.

20. Page 8, lines 30-33: Two issues regarding your thersholds for coloring MSESS- and CB-values in the tabel:

20.1. Your thersholds in the text partly don't match the thresholds mentioned in the caption. Please check carefully!

Answer: We clarified the captions for Tables 1-3 and changed the colouring slightly to provide a clear and consistent description (see also answers to 20.2 and 21).

20.2. The choice of these thresholds seems totally arbitrary here. Is there some though behind it? If so, please explain. At least for the MSESS it would have made much more sense to me to also use a certain significance threshold as justification for coloring here.

Answer: The choice of the thresholds in Tables 1-3 is based on the typical level, above which the skill scores are regarded significant by the bootstrapping. This is about 0.3 for the MSESS and about 0.4 for the correlation (highlighted in green). For the conditional bias it was in a range between +/-0.2. Negative MSESS and correlation values are highlighted in red, as well as CB values beyond +/-0.3. Values in between are regarded as not significant and therefore not marked. We have clarified this in the revised manuscript.

21. Page 9, line 1-3: Why do you suddenly use a t-test for assessing statistical significance? It would have been possible to use the same bootstrapping approach for CB_AV, too.

Answer: We agree that this was confusing. The conditional bias has an optimal value of zero. Therefore, the 0-hypothesys is different. To be consistent with the other scores, we now apply a common threshold of 5% (+/-0.05) to indicate a distinct difference of the skill scores between global and regional ensemble (added value), as well based on the typical level above which the differences are significant. Differences above 0.05 are now marked in green, below -0.05 in red and between +/-0.05 in white. We have changed the text and the captions accordingly.

22. Page 10, line 4-5: As mentioned above, it is no open question if there is a dependency. This would be one of the instances where I would ask you for a reformulation in the sense of "demonstrating skill bias dependency" (see my general comment 2 above).

Answer: We have rephrased this sentence and the remainder of this section according to comment 2.

23. Page 10, lines 8 and follwing, as well as Fig. 5: I think it might be quite misleading to demonstrate this bias dependency compared to the unitialzed simulations. They suffer from a bias, too, and currently it is not clear from the manuscript whether you reduce the number of unitialised simulations for the reference forecast, too, or use their full ensemble in every instance. I strongly suggest to demontrate the skill bias comparing to the climatology as reference forecast. This one is unbiased!

Answer: Originally we have exclusively focused on the uninitialized historicals as reference dataset, as is mostly done in the decadal forecast research community (see also references in our manuscript). This is motivated as follows: for future projections of the upcoming decades the RCP scenarios are used. When analysing past decades, the analogue to these RCP scenarios are the uninitialized historicals. Hence, a decadal prediction system is regarded as skilful, when the initialised hindcasts are closer to the observations than the uninitialized historicals, and these historicals are therefore used as reference. However, in one of the former revisions of our manuscript we additionally included the climatology as reference as suggested by one of the Reviewers, since we agreed with this Reviewer that this may improve the scientific value of our study. Nevertheless, our main focus still lies on the uninitialized historicals. And since we always use the same full ensemble of 10 members for the uninitialized historicals in Figure 5, we think that it is suitable for our purposes to demonstrate the improvement of the prediction skill when the number of ensemble members is increased. We have added this information in section 4.3. Therefore, depending on the Editors decision we would prefer to keep Figure 5 as it is in the revised manuscript.

24. Page 11, lines 3-7: I think you definitely should remind the reader here once again that your box-whisker plots do not contain the full uncertainty of the skill score estimates!

Answer: We are not sure if we fully got the intention of the remark, beyond what was already changed w.r.t to remark 23. Nevertheless, we added a remark in the caption of figure 5 to indicate that it just covers the uncertainty of the skill estimates due to the sample size.

25. Page 11, lines 11-12: This is also included in the CMIP6-DCPP requirements/recommendations,sp please include a reference to Boer et al. (2016) here, too.

Answer: We have included the reference to the DCPP and to Boer et al. (2016) in the revised manuscript.

26. Page 11, line 14: Please rephrase the research question and your following text in the sense of my general comment 3, so make it mor clear that you are addressing the issue of few initializations here.

Answer: We have rephrased the research question as suggested by the Reviewer and reformulated the following text accordingly.

27. Page 11, line 23: Replace "starting years" by "initializations only".

Answer: We have changed it to “initializations” in the revised manuscript.

28. Page 11, line 24: Replace "starting years" by "initializations".

Answer: We have changed it to “initializations” in the revised manuscript.

29. Page 12, line 5: Replace "predictability in" by "prediction skill of".

Answer: We have rephrased it as suggested.

30. Page 12, line 27-28: Rephrase. Maybe something like "Based on MPI_b1 data [...] we could show that results derived from only those five initializations used in our study qualitatively agree with results based a full set of annual initializations".

Answer: We have rephrased it to “Based on the MPI_b1 data, it was shown that results derived from only five initializations used in this study qualitatively agree with results based on the full set of annual initializations.”

31. Page 12, line 32: Replace "predictability" by "prediction skill".

Answer: We have changed it as suggested.

32. Page 12, line 19: Be cautious with such statements regarding the relevance of drifts (and subsequent corrections) in case of anomaly initializations. Some studies show that these feature drifts, too. Reformulate to something like "the general expectation is that drift correction is less important for prediction systems employing anomaly initialization".

Answer: We thank the Reviewer for this hint. We have changed it accordingly.

33. Page 12, line 22: Replace "starting dates" by "initializations".

Answer: We have changed it to “initializations”.

34. Page 14, line 1: Replace "predictability" by "prediction skill".

Answer: We have changed it as suggested.

35. Page 14, line 3: Maybe it's worth mentioning that this huge amount of 1000 RCM model years may also be a valuable set for other studies, not necessarily related to decadal prediction. I mean, it's a huge dataset in comparably high resolution representing the European climate of the recent past…

Answer: We agree that such a sample is very valuable not only for decadal predictions but also beyond. We have added this information in the revised version and named return periods of extreme events as example.