J Day et al. Response to reviewers comments on “The Arctic Predictability and Prediction on Seasonal-to-Interannual TimEscales (APPOSITE) data set”

We would like to thank the reviewers for taking the time to carefully read this paper and for some very useful suggestions. Whilst we agree that this dataset is ideal for some of the additional analyses suggested by the reviewers and that these would be very informative. As the APPOSITE project has come to an end, we would like to point out that the primary role of this manuscript is to provide a descriptive reference for this dataset, so that it is well described for future use. Therefore our primary action in response to the comments has been to clarify and expand on the description of the experiment and archived data, where suggested by each reviewer. That said, we have taken the time to follow a suggestion by both reviewers to examine the initial state dependence of sea ice predictability and have included a new subsection and additional figure on this point.

Reviewer 1:

In this paper a multi-model protocol for analysing potential model predictability is introduced, focusing on the potential predictability of the Arctic sea ice conditions on the seasonal to interannual timescale. The setup of the ensemble simulations is explained as well as the diagnostics used to analyse potential predictability of Arctic sea ice extent and volume. Seven different models have contributed to create a dataset following the basic guidelines of this protocol, with some difference in the more specific details such as ensemble size and number of ensemble start dates. The results for the ensembles of four of these models regarding potential Arctic sea ice predictability have previously been discussed in a paper by Tietsche et al. (2014), while the results for the remaining three models are added to the discussion for this paper.

In general I appreciate the effort of the authors to make the data available to the broader scientific community and to use this publication as a reference for the setup of the experiment protocol. Analysing potential predictability and the differences therein between GCMs is certainly an important area of research, especially as a tool to inform seasonal prediction systems of the feasibility of future improvements. The paper is generally well written and the structure is straight forward. While I appreciate the authors’ choice to keep this publication short and concise, I do have some comments that might increase the length of the paper quite a bit. My main point of critique is that the paper is very close to the previous publication by Tietsche et al. (2014) without presenting a more detailed description of the experimental setup, and without discussing the new results equally detailed as the previous study. Since both aspects are the main points of this paper, they should be extended, still keeping them as separate aspects of the same publication, i.e. first the discussion of the protocol, then the application to the newly contributed models, highlighting the importance of both.

General comments

As a first comment and to repeat my question of the summary, could the authors be more specific regarding the focus of this paper and how it differs from the Tietsche et al. (2014) publication. I assume you want to equally focus on the results for the additional three models as well as on the general setup of the protocol. But at the moment I would claim that both parts are a bit too short and not very detailed.

In the Abstract and Introduction, we have been more specific about the goals of the analysis, which is to provide an updated estimate the predictability forecast horizon for sea-ice extent and volume also mentioning the additional work on sea ice extent and volume predictability initial state dependence.

Some more specific examples regarding the experimental setup:
When you write about the high, low and medium sea ice states used for initialisation, how is that reflected in the actual ensemble start dates? Does this relate to the sea ice volume, the sea ice area or average sea ice thickness? Are they separated in some way in the archiving structure? Are you trying to estimate the impacts of different initial conditions by this approach, even though some models only have 8 different start dates, which would make it difficult to actually assess differences in the predictability caused by the initial state?
Choosing the start dates was essentially left up to the participating group, but we encouraged them to sample a range of initial states based on pan-arctic extent and volume. The aim was to investigate state dependence of sea ice predictability. These points are made explicit in Sec 2.2.

When you say "well spaced" (page 8815, line 18) how is this defined? Was there a minimum spacing between successive start dates that you have generally defined for all models to insure independence of the initial state?
As the modelling centres chose their own start dates there is a bit of a range, the minimum spacing is 3 years for GFDL, but longer for other models.

How was the length of the control run defined? Different models have different spin up times and might take longer to equilibrate. After only 100 years I wouldn't think any model has really equilibrated, as can be seen by the strong drift of most of the models.
Could you comment on some of these details, stating advantages and disadvantages of the choices you had to make to generate this dataset. Also, in this context, the time axis for the panels in Figure 1 doesn't make much sense to me. The start date of each model control seems more or less random, even though the text reads they started from (the same?) static state oceanic depth profile.
Again, the particulars of initialisation, spinup and length of control were dependant on the modelling groups. Some groups had a 1990/Present day simulation with their model, but many did not. As this is not part of the CMIP5 DECK so groups either had to create the necessary boundary conditions and start a fresh simulation. As no groups outside Reading were funded to do these simulations it was difficult to standardise this approach. However, since every model has been spunup for at least 100 years, intermodal differences in climatology are unlikely to be affected significantly by these differences. Since we are looking at initial value predictability only over the first three years, it is unlikely that issues such as drift play a large role in the assessment of predictability. We have expanded the text in this section to make this more apparent.

It's worth noting that even after 1000s of years, many climate models still drift, so this is something we have to live with. In practice, even with 200 years of model time series, models with pronounced low-frequency variability can exhibit apparent drifts even when they are in "equilibrium" purely due to the particular phases of variability captured in the window used for trend analysis.

Effectively the times in Figure 1 are random since this is just the model clock year in the control run. We only show the period of the model that was used to calculate the reference climate mean and *standard deviation. The spinup period of the models was not collected from the centres or archived. We have made these points explicit in the Figure 1 caption and Section 2.1 text.*

Were the SST perturbations applied globally, also in areas of sea ice cover?
Yes, we make it explicit that they were applied at all ocean cells.

Regarding the two metrics, were they applied to detrended monthly means? If so, was the detrending based on the control or all ensemble members? It would simplify the explanations for the metrics if you would actually expand the expectation value as was done in Collins (2002), also to show which normalization you chose (what is sigma?).
This is the standard deviation of the model climate, as shown in Figure 4. We have made this clearer in the text.

What kind of significance test was applied to the ACC?
We used a T-test, details are now given in the text.

Are there any specific plans to extend this dataset, i.e. to include more models? Or to use this dataset for other predictability studies?
There are plans for this, but they are dependent on the outcome of funding proposals. We think these plans are too tentative to be worth mentioning in the text.

Some more specific examples regarding the results:

The sea ice models in this study differ in many aspects. Could you comment a bit on how this affects the results? For example, do models with similar albedo and melt pond parametrizations produce similar results, or do models with similar sea ice dynamics (number of sea ice classes and so on) produce similar mean states and climate variability? I know this is a difficult questions, since the other model components show significant differences as well. However, it would be interesting to know whether some systematic differences can be identified.
As the reviewer states, this is a difficult question to answer. All we can say is that we have not identified any such links between sea ice model formulation and other properties. We believe a more targeted experiment would be required to say more about this.

Could you please expand the paragraph about the mean state and climate variability. For one, it is not surprising that the mean states of the models are different compared to the mean state of the observations, which have been recorded over a shorter period of time and under transient forcing conditions.

We have added to the discussion here. However, because this is simply designed to highlight the variety in model climate states rather than robustly assess the realism of each model, we do not present a detailed assessment of model climate. This aim is also made explicit in the text.

Furthermore, could you comment on how model variability and mean state affect the predictability metrics.

This is an important question, however with the number of models available we only have 7 data points to derive any relationships, which we believe is too few to do anything robustly. We are planning to extend this dataset as part of a later proposal and come back to this point. We also include this as an open question in the conclusions section.

What are the consequences of the different drifts in the models? Do you expect a more equilibrated model to provide a more accurate estimate of potential predictability?

We have taken account of this in the metrics used by using a time varying climatology in the case of ACC. This is explained in more detail in the text.

Why didn't you apply any of the spatial predictability metrics which were used by Tietsche et al. (2014)? What about the other start dates provided, especially January? Since the extended results of this paper are mentioned as one of the two major contributions of this study, it would be nice if the paragraphs about the model results (page 8818) could be expanded, providing more details on the differences and similarities in predictability between the models and possible reasons for that.

We have added a paragraph to the end of Section 3.2 to discuss some of the open questions relating the predictability to climate and some potential next steps. We have also clarified that we are extending the analysis of Tietsche et al. in particular to assess the limit of extent and volume predictability from July. Hence we do not utilise the Jan predictions, or the spatial measures.

Page 8818, lines 12-15: How does this relate to the results of the current study?

Have added 'Indicating that the winter sea ice extent predictability horizon may be significantly beyond the 3 years simulated in these experiments' to the end of this sentence.

Page 8818, line 23: There is always a chance that you remove internal variability by detrending, also for a longer timeseries. It is just less likely.

Have added "is likely to significantly", the point being that it will be enough to significantly affect the predictability metric."

Page 8818, lines 26-27, and page 8819, lines 1-3: This paragraph is difficult to read. Maybe you could break up the sentences.

This paragraph has been rewritten.

Page 8819, lines 6-7: The differences of the mean state and variability between models and observations wasn't discussed in any detail.

I have changed this to say we have presented the mean state and variability.

Page 8819, line 17: Not really true for E6F (early loss of predictability for sea ice volume; no re-emergence of predictability for NRMSE).

This statement is less true for E6F, we have changed this "Sea ice volume is ***generally*** *more predictable than sea ice extent"*

Minor comments:

Page 8811, line 16: Change to “Unprecedented”, “opportunities”, “businesses”.

Page 8811, line 17: Change to “but has also”.

Page 8811, line 23: “appreciation”.

All above changed

Page 8812, line 1: What do you mean by “significantly skillful”? Could you also give a reference here?

Changed to “have statistically significant skill”

Page 8812, lines 9-11: Please rephrase this sentence. Be more specific about this “fundamental limit”, which has different timescales for the atmosphere and the sea ice.

Done

Page 8812, lines 20-21: Please expand this. What are the disadvantaged of potential predictability studies? How does model uncertainty affect predictability estimates?

We have added some additional discussion here.

Page 8813, line 5: Change to “: : : climate variables as well. In order: : :”.

Changed as suggested

Page 8813, line 10: Differences in design such as?

Page 8813, line 12: Differences in the results such as?

Have rewritten this section on motivations.

Page 8813, lines 13-16: Again, could you name some of the differences, either here or before?

OK

Page 8814, line 22: Change to “sea ice”.

Done

Page 8815, line 1: Change to “distribution, as well as”.

Done

Page 8815, lines 11-13: Can you quantify this/be more specific? Does this have consequences for summer sea ice predictability when it comes to different model mean states?

Added as stated above

Page 8815, line 20: Change to “depending on”.

Done

Page 8816, line 8: Remove comma at the end.

Done

Page 8816, line 21: Change to “inter-model”.

Done

Page 8818: Mention Figure 5 again, after first sentence of 3.2 and 3.3.
Page 8819, line 14: Change to “interannual”.
Page 8820, line 7: Change to “constraints:”.
Page 8820, lines 8-11: Could you give a reference here?
Page 8820, line 23: Change to “submodel&frequency”.
Page 8820, line 23 onwards: Check for text size and font here and on the next page.
Page 8820, line 25: Is it “1” (this line) or “r1” (next page, line 1).
Figure 2 and 3: Is the average taken over the entire simulation length or only for the years after the spin-up?
Figure 4: Mention detrending in caption.

All Done as suggested.

Anonymous Referee #2

The manuscript presents an updated version of the APPOSITE dataset that is originally presented and discussed in Tietsche et al 2014 and Day et al 2014. In its current version, the manuscript adds unfortunately little new information or insights into sea ice predictability to these two papers, and I feel it is, as it stands, a missed chance to use the dataset to explore issues that are at present topical in the field. I would encourage the authors to extend their analysis.

Since publication of Tietsche et al. (2014), the APPOSITE protocol was followed by a number of additional models and this database has been made openly available as a community resource. This is why we believe that it is useful to publish an extend the description of the dataset and update the results of Tietsche et al. We agree that there are still many open questions in this area, which is why we have made the effort to make this data openly available. It provides a unique resource to investigate initial value predictability in multiple models.

I suggest below a few ideas to explore. How does predictability depend on mean state? The APPOSITE dataset, with its start dates split between high,medium, and low initial conditions (p8815 L18), is currently the best opportunity to explore this question. If you find that the number of ensembles/runs is still not large enough to yield statistically robust results, this finding would still be useful for the community - I suspect the answer will depend on whether in fact there are (meaningful) inherent differences in predictability with mean state. Given current trends in sea ice in observations, exploring this issue is key.

How can we understand the inter-model differences in predictability? While the patterns in change of predictability with time are similar across models (e.g., predictability barrier in SIV in early summer, winter>summer SIE predictability in years 2,3), there is a considerable spread in predictability across models as you point out in the conclusions (as an aside, I would guess given your ensemble size that the inter-model differences are significant, but it would be good to calculate and show this). This is a significant result. I note that in Day et al 2014 (Jclim), you explore links between predictability and persistence, and persistence and mean state. It would be good to do this with the current larger dataset. Are models with higher predictability more ’persistent’ (Figure

1 shows models have varying degrees of persistence in their control runs)? It has been shown (B-W and Bitz, 2014) that models with thicker sea ice tend to have longer thickness persistence timescales - does this help explain inter-model differences? By looking at Figure 4 and 5, it's hard to figure out if there's a link between total volume and predictability. Perhaps a scatter plot of e.g., mean NRMSE over year 1 against mean SIV would help. (You could even split each model into its 3 high/medium/low ICs and obtain 6*3 datapoints).
We agree that the question of how predictability depends on model mean state, or other properties of model climate is a crucial one. However, we feel that given the limited set of models it will be difficult to infer any robust relationships. However as part of a follow-up proposal we intend to extend these runs to other models so that such an analysis will be possible.

We have however extended our analysis in this work to investigate how initial value predictability depends on whether the model is in a high, medium or low state at its initial state. This is in a separate section of Section 3 (3.4).

Can you extend the dynamic v thermodynamic analysis of Tietsche et al 2014 (see their section 3.3, Fig3) to more models? Discerning which physical process leads to loss of predictability, particularly at seasonal timescales, would be an important result. Additionally, considering if the relative importance of different processes varies between different initialization seasons (January vs July) would be equally insightful.

Unfortunately the diagnostics required to perform this analysis were not available for models other than MPI and HadGEM.

Minor:

There are several spelling mistakes - please proof read cautiously

We have thoroughly proofread the document and removed a number of spelling mistakes.