Comments Reviewer #1

I would like to thank the authors for considering my comments and changing the manuscript accordingly. In my opinion the manuscript has been improved by the more focused discussion of the experimental setup and the inclusion of the analysis on initial state dependence of potential predictability. The latter specifically gives an idea of what kind of analysis is possible with this dataset that hasn't been done yet.

However, since you have decided to focus on the experimental setup of the protocol there is still one major aspect I would like the authors to comment on.

Could you include one sub-section as part of an outlook which discusses how this protocol could be improved and what is missing at this stage? This could be used as a guideline for future experimental setups. Since you gathered and analysed this dataset you have a good insight into what was missing and what kind of analysis you would have liked to do but which you couldn't because of the restrictive nature of this (and any other) protocol.

Therefore, could you, as part of a kind of overview/summary/outlook sub-section,

-Highlight the disadvantages of the current datasets, e.g. number of ensemble members, start dates, (especially with regards to minimum, average and maximum extent/volume years), output (such as tendencies necessary for detailed analysis), start times, control setup, forcing, simulation length and/or other aspects.

-Describe changes to the protocol which should be applied in the future, e.g., clearly defined time intervals between start years, number of ensemble members and start dates necessary to allow for robust and statistically significant diagnostics, number of participating models and so on.

-Discuss briefly which parts of the protocol are most important in terms of enforcing a common setup between participating models. Is it number of start dates rather than members, integration length of the ensemble members, the length of the control simulation etc.

I know these aspects and their importance differ depending on what part of the climate system you are looking at but that is precisely why you should comment on this from the perspective of potential interannual sea ice predictability. Do you expect differences in these aspects when looking at Antarctic sea ice?

In this context it would also be good to have a document (Supplementary material? ) explaining the minimum experimental setup to take part in APPOSITE, including such technical aspects as output variables and frequency, simulation length, ensemble members and so on. But maybe you have already supplied this to the British Atmospheric Data Centre.

Other minor comments:

Line 139: Change "the determining" to "determining"

Line 181: Change to "uses"

Line 218: Change to "timeseries"

Line 262: Move "CanCM" to "simulations"

Line 263: Add "had been run for a longer control period in the fixed" or something similar

Lines 291-293: There is something missing after “model states of the”

Line 293: Delete “is”

Line 297: Add “in”

There are two minor things I mentioned in the last review that you haven’t changed:

Line 371: Check for text size and font here and onwards

Line 368 and 371: Is it “1” or “r1” for “<run>” in the control case

Figure 6 caption: Change “os” to “of”

Figure 6 caption: What is the dashed line (average)

Figure 6: Maybe explicitly mention in the text/caption that a significance test with so little independent data points as used for Figure 6 doesn’t make much sense and therefore this result can only be seen as an indication.

Figure 6: Please add the number of start years for each ensemble and each case (low, medium, high), either in the caption, table 1 or in brackets in the figure legend.