I enjoyed reading this article very much. The results are relevant for the scientific community, numerical weather prediction centers, and forecast users. The forecast verification process is supported by a solid and sophisticated methodological base, and the forecast improvements and deficiencies are honestly highlighted without over or understating the findings. Furthermore, the manuscript is well written, and the figures illustrate the outcome of the study appropriately. I include below a few remarks and suggestions, which are mostly minor, and I hope that these will help the authors in the revision process. I recommend the publication of with manuscript once those (minor) points are addressed.

GENERAL COMMENTS

The Introduction and Method sections provide a very good overview of the system. However, I think some details on the probabilistic nature of the forecast are missing. Underlining the higher compatibility of a coupled model configuration with the ECMWF ensemble forecast system (i.e. the ice can evolve independently in each ensemble member, unlike in the persistence based strategy) would represent a nice addition to the study.

What about the melting season? I expect the impact of the sea ice on the ocean and land weather to be limited in summer because of the milder temperature gradients and winds. However, the demand for good ice forecasts might peak during this season. I would briefly mention whether the features of the dynamical system are appropriate also for the summer months. I also think a brief reference to what happens to the Southern Ocean sea ice might be appropriate.

The study focuses on a single winter season. Given the large number of forecasts
analyzed, I expect the results to be solid. However, I think it might be good adding a characterization of the sea ice state during that winter in comparison to the climatological state, and discussing whether the results might be sensitive/influenced by potentially anomalous conditions (e.g. fast ice drift, abrupt melting events, etc.)

Figure 4 clearly shows that OCEAN5 reanalysis is biased, and you describe this well in the text. However, I think giving some more context on the origin of this bias would be helpful for the readers.

The fact that the thickness is not coupled implies that the thermodynamical transition at the ice edge is probably not well simulated also by the current dynamical system. Could you quantify the impact of this on the evolution of the internal boundary layer? Is the effect of a progressively reduced thickness towards the marginal ice zone negligible compared to the reduction in concentration? I expect this would also changes with the progressing of the freezing season. I think some more details on this in the discussion/conclusion section would be interesting for the reader.

I would like to point out that the using the AMSR2 derived sea ice concentration has also some drawbacks. It certainly comes with a desirable higher resolution because it uses higher frequencies. However, the effect of clouds on the microwave signal at higher frequencies is substantial and can penalize the quality of the retrieval, particularly across the marginal ice zone where clouds are not uncommon.

SPECIFIC COMMENTS

Line 71: I would not consider obs-SSTSIC a real forecast but rather an hindcast or an AMIP type simulation.

Line 75: I think it is worth mentioning here that the sea ice description of OSTIA comes from OSI-SAF. You mention this later in the result section, but I think stating this here would help the reader to understand the verification method.

Line 94: I suggest expressing the typical Arctic resolution of the ORCA025 grid also in km.

Line 96: Do you mean “...that is coupled to the atmosphere is...“?

FIGURES
Fig. 1: I suggest using a different colormap with the white color centered on zero. I don’t like seeing the rectangular domain of the polar stereographic grid in pink. Using red and green for the boxes might not be color friendly.

Fig. 2: Labels and titles are too small. I think it is ok to lose the outlier point in the OSI-SAF timeseries (plots a and c), probably caused by a partial observational coverage on that day.

Fig. 4 and 6: Labels are too small, and I suggest using the scientific notation to improve the readability of the plots.