Dear Referee,

Thank you for the time that you have spent on our manuscript. We are happy with your positive response and grateful for your comments and suggestions. These certainly contributed to improving the quality of our manuscript.

Below you will find a summary of the changes that we have made throughout the manuscript to address all of your suggestions. The replies to your comments are written in blue, while your comments are reproduced in black. Please, notice that line, page, and figure numbers mentioned in our rebuttal letter refer to the new version of the manuscript unless stated otherwise.

Yours sincerely and on behalf of all co-authors,

Leandro Ponsoni

Anonymous Referee #1

GENERAL OVERVIEW

The manuscript presents a statistical model for predicting the pan-Arctic Sea Ice Volume (SIV) anomaly on an interannual timescale. The long-term variability and the seasonal cycle have been subtracted to focus on the interannual SIV anomalies only, therefore excluding other better-understood signals. The statistical model is trained on the output of three coupled climate models produced in the frame of the HighResMIP. A low and high-resolution version of each model is analyzed.

The first part of the study inspects the capability of seven predictors to represent the sea ice volume up to 12 months in advance. The authors focus on two target months: March (post-winter conditions) and September (late summer conditions). These predictors are tested and combined, both on a pan-Arctic and regional scale. The results show that the best predictive skill comes from the SIV itself, and by the Sea Ice Thickness (SIT), while the other considered variables are progressively less skillful.

The study presents afterward a method to determine some optimal locations that are representative of the SIV anomaly variance. Those locations are picked in a smart way to avoid clustering of points in certain regions, while other parts of the Arctic Ocean are underrepresented. The authors show that the statistical model can reconstruct approximately 70% of the SIV anomaly variance when fed with only 4 well-placed locations.

Even though the results here presented are in line with our expectations and not surprising, the manuscript tries to establish a robust protocol to predict the SIV anomaly. Furthermore, the fact that a large part of this variance can be predicted with only a few sparse observations in strategic locations is certainly interesting and can guide the design of future observation campaign in the Arctic region. The comparison of high and low resolutions contributes to the ongoing discussion in the modeling community about the benefit of resolving small features compared to the computational costs.

The approach followed by the authors as well as the application of this methodology to the SIV anomaly is quite novel. The purpose of the work is well presented and the methodology is adequately

explained. The model data here analyzed are cutting edge in terms of model physics and resolution. The manuscript is well written and the figures and tables convey the message effectively.

The content of the study is certainly appropriate for The Cryosphere and I recommend the publication of this manuscript. Below I include a few minor points and suggestions that the authors should be able to address easily.

Again, we thank the referee for her/his time and detailed revision of our manuscript. We appreciated very much her/his comments, which were all taken into account in the revised version of the paper. Below, we answer point-by-point all specific comments.

SPECIFIC COMMENTS

The manuscript provides several sampling locations with a multi-model approach. In my understanding, these locations are computed based on annually-averaged fields. I am wondering if the sampling locations could be different for different target months. Also, some of the selected sampling locations might be ice-free in some periods of the year. Could the authors comment on this?
All results of the manuscripts are based on monthly-averaged fields. This point is clarified in the text [pg. 4, l.18–19].

Except from Sec. 3.1, where we first assessed the performance of different predictors by focusing on March and September (Figs. 4 and 5), the other sections do not make a distinction of months.

However, we understand the referee's comments since we had posed the same question to ourselves during the preparation of the manuscript's first version. We have decided to avoid the distinction of months for the following reasons:

i. The motivation of the manuscript is to provide support for a **year-round *in situ* monitoring system**. Thus, those are sampling locations that better reproduce/predict the pan-Arctic sea ice volume taking into account continuous monitoring throughout the entire year.

ii. A distinction of months would likely suggest relatively different locations. In the real world, **this would require a re-positioning of observational platforms** (e.g., moorings and/or buoys) every month.

iii. The fact that some sampling locations might be ice-free in some periods of the year **is part of the time-series variability and it brings predictability to the statistical model as well**. If the grid-point is ice-free for long periods, predictors as SIT, SIC and Drift will be disregarded by the correlation map criterion. The SST predictor can still be useful even from grid-cells which are mostly of the year ice-free. Nevertheless, the four most performance locations are likely covered by sea ice during the entire year, for most of the years.

iv. By splitting the time series into 12 parts, we substantially **reduce the number of points for training and applying the statistical model**. The fact that the statistical model is randomly trained (70% of the data) and applied (30% of the data) within a Monte Carlo (MC) scheme (500 reproductions) give us statistical robustness to assume that this configuration is the best scenario for a year-round sampling system. We have tried to increase the number of MC interactions but it turned out that 500 is already a safe threshold.

I believe that an interesting exercise would be comparing the performance of the statistical model in the optimal location to that in randomly chosen locations. This would show that the described method is robust and in fact, needed.
We absolutely agree, thanks for the interesting suggestion. To address it, we have compared the RMSE and R2 calculated between the original and our-methodology-based reconstructed SIV anomalies (as shown in Fig. 11a,b) against the same two metrics estimated by randomly chosen locations. To do so, we have determined 100 combinations of 10 randomly chosen locations. For each combination, we reconstructed the SIV anomaly using data from the 1st location, the 1st–2nd, the 1st–3rd, the 1st–4th, and so on. Fig. A (this rebuttal letter) shows 2 of the 100 sets of randomly chosen locations. For the sake of fairness, we have used only predictors from grid points enclosed into the region highlighted by the red line in Fig. A. This region represents our global region of influence as defined by Fig. 8 (now this line is also plotted in Fig. 9). It is worthwhile saying that 100 combinations of randomly locations already provide robust statistics for such a comparison.

Fig. B shows that the SIV reconstructions based on our methodology (and optimally selected locations) are more skillful compared to the predictions provided by the randomly chosen locations, taking into account both metrics (RMSE and R2). This is valid for all models, considering a single location and/or any combination of 1 up to 10 locations.

These results, and respective supporting Fig. B, were incorporated into the new version of the manuscript (pg. 23, 1–9, Fig. 12).

While the current model results provide an average representation of some variables inside a grid cell with a substantial extension, and the gradients between different cells are generally small, real-world observations would be much more localized and heterogeneous. Would this heterogeneity introduce some sampling errors and consequently require more observations to explain the SIV anomaly variance?
As the referee highlighted, real-world observations are much more heterogeneous than averaged grid cell values. Compared to other oceanographic parameters such as temperature and salinity (unless in regions marked by steep frontal systems), this issue is even more pronounced when looking at the sea ice due to its inherent roughness. Thus, we indeed expect that this heterogeneity may be a source of uncertainties in a real observing system. We also agree that more observations could attenuate these uncertainties. This is a very important point that was quickly addressed in the first manuscript's version [former pg. 24, lines 3–4]. We have added a few more words to make this point clear in the manuscript [pg. 26, l. 14–19].

Is the whole time period (∼150 years) necessary to reach the described results? I think it would be interesting to assess how many years of observations would be necessary to train adequately the statistical model here presented, and robustly reproduce the HighResMIP results.
We have used model outputs from coupled historical runs, referred to as "hist-1950", performed within the context of HighResMIP. So, from all model configurations the data spans about 65 years, starting in the early 1950s and finishing in mid-2010s [pg. 4, l. 6–8]. We understand that using these 65-years is indeed necessary to achieve statistical robustness.

1 – Line 16: It is worth mentioning also the SMOS sea ice thickness product.
SMOS is now mentioned in the text [pg. 3, l. 16].

2.1 – Line 6: Are the analysis on AWI-CM performed on the original FESOM2 grid or was the model output interpolated to a regular grid?

Sea-ice concentration (SIC) was provided by the AWI group on the original atmosphere grid in the framework of the PRIMAVERA project. **Sea-ice thickness (SIT)**, **sea-surface temperature (SST)** and **sea-ice drift speed (Drift)** were also provided by AWI but on a 1-degree regular grid also in the framework of the PRIMAVERA project. **Sea-ice area (SIA)** was computed from the SIC files and the atmosphere grid-cell area, while **Sea-ice volume (SIV)** was computed from the SIT files and the ocean grid-cell area (Docquier et al., 2019). Finally, ocean heat transport (OHT) was computed by the AWI group directly from the raw data. Additional information is presented in Section 2.1 of Docquier et al. (2019).

2.1 – Line 7: I would mention that the resolution difference between HR and LR in the Arctic is much lower in AWI-CM compared to the other two systems.

This is indeed a good point. We thank the reviewer for spotting that. While the ocean resolution in AWI-LR and AWI-HR varies between 24 and 110km, and between 10 and 60km, respectively (with higher resolution in dynamically active regions), the ocean resolution is almost similar in the Arctic Ocean (~25km). We brought this information to the text [pg. 4, l. 33–34 to pg. 5, l.1–4]. In addition, since the grid used by AWI is not trivial to understand without a supporting plot, we are directing the reader to Fig. 4 of Sein et al. (2016).

2.2 – Line 34: Is there a particular reason for choosing AWI-LR?

No, there isn't a particular reason for choosing AWI-LR. We selected AWI-LR as the example-case. "AWI" is the first model in our alphabetically-sorted list and, in the other model-comparative figures (e.g., Fig. 6), we always referred first to the low resolution (LR) version. We clarified this point in Fig. 2's caption.

2.4 – Line 12: Be specific about the "common grid". Is it a low or high-resolution grid. Can this have an impact on the results?

We agree that this point requires clarification. As suggested by the score maps in Fig. 6, each model configuration indicates its own best sampling location (smallest RMSE in the score map). However, the RMSE values show that overall there is a good agreement on the regions with high scores (small RMSE values represented by yellow shades). To achieve an ensemble best location we first applied Eq. 4 to normalize all score maps between 0 and 1 so that the models have the same weight in the averaging step (Fig. C, first column). However, since the models have different grid-resolution, we have interpolated the score maps from the different models into a common 1°×1° grid. By performing this step, we can calculate an ensemble mean score map.

The interpolation of the individual score maps into a common 1°×1° grid for further computation of an ensemble mean score map has no impact on the results [pg. 9, l. 32–33]**.**

Notice that the interpolated score maps (Fig. C, second column) preserve the best performance regions.

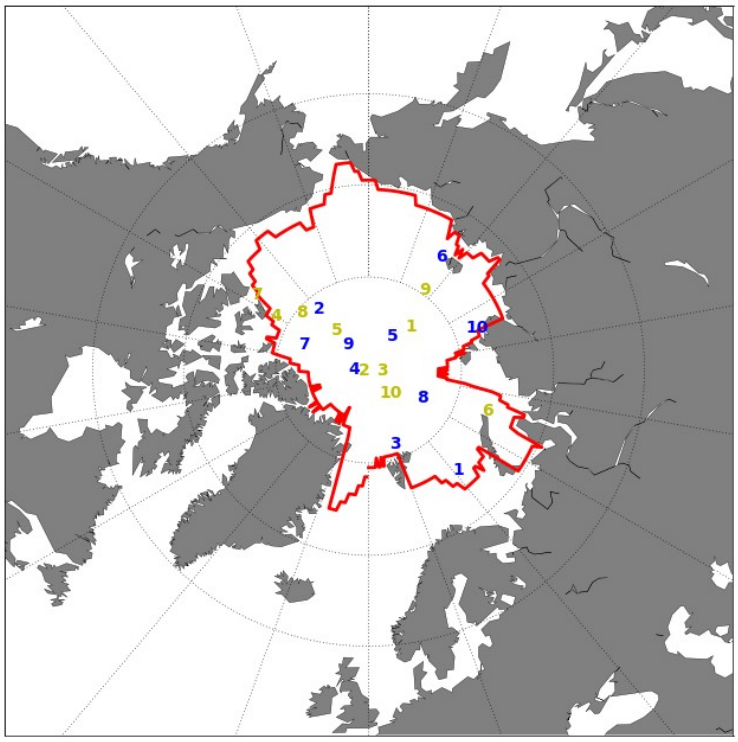

Fig. A: Map displaying two examples (out of 100) of randomly chosen locations. All random locations are placed into the area enclosed by the red line. This region represents our global region of influence as defined in Fig. 8.

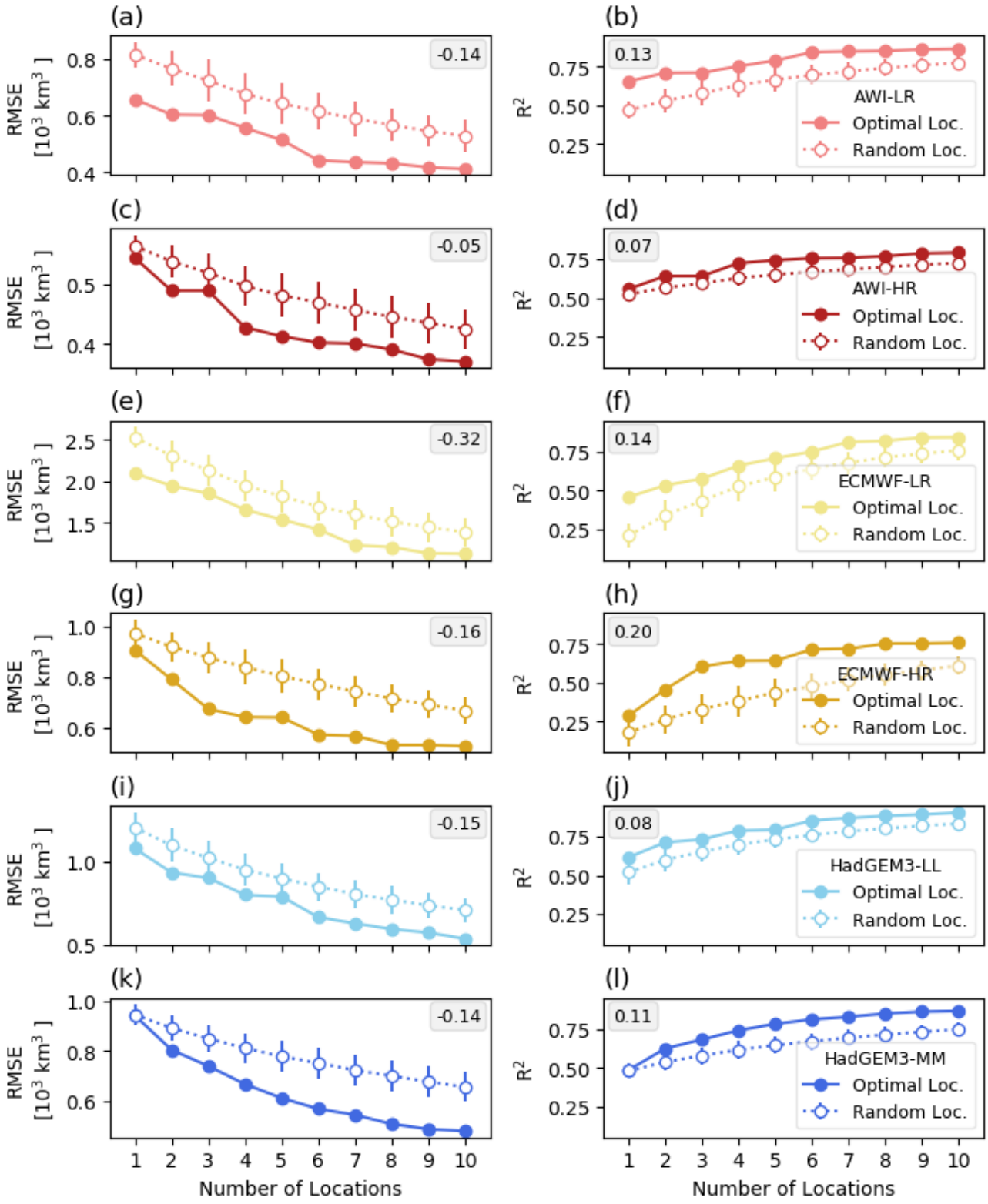

Fig. B: Root Mean Squared Error (RMSE; left column) and coefficient of determination (R^2; right column) calculated between the original and reconstructed SIV anomalies. The reconstructed SIV volume anomalies are based on the optimally selected locations following our methodology (full dots), as well as by randomly chosen locations (empty dots). In the last case, 100 sets of 10 randomly chosen locations are used. For each of the 100 sets, the SIV anomaly is reconstructed using data from the 1st location, the 1st–2nd, the 1st–3rd, ..., the 1st–10th. The random locations are all placed into the region enclosed by the red line shown in Fig. A. The vertical bars associated with the empty dots represent the one standard deviation.

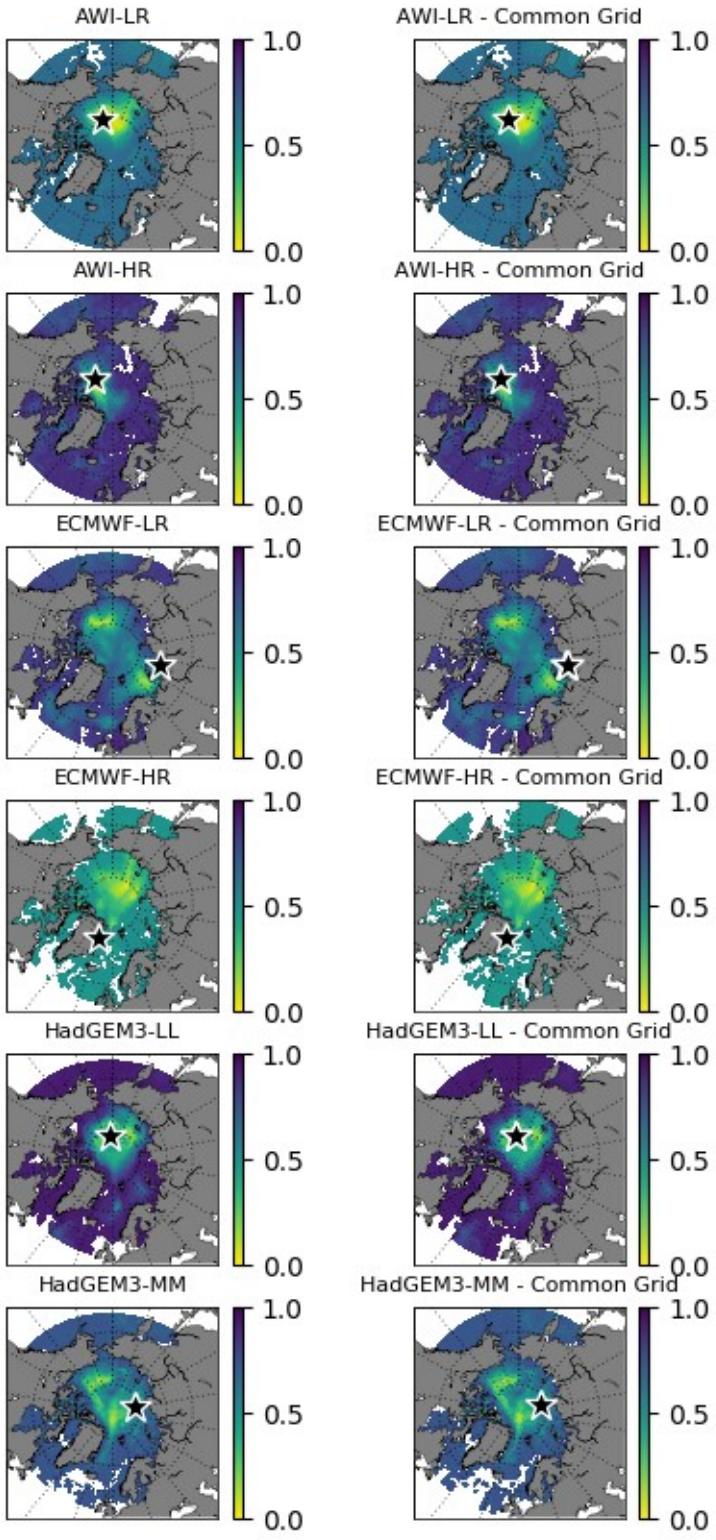

Fig. C: Normalized score maps calculated for the different model outputs with the original grid (left column) and after the interpolation to the common 1°×1° grid (right column). Notice that the interpolation has no impact on the best performance regions (shades of yellow). The interpolation is a required step for calculating an ensemble score map since the models have different resolutions.