Dear Editor,

Thank you for the time that you have spent on our manuscript. We are happy with your feedback and grateful for your comments and suggestions. In addition to the previously answered referee's comments, 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**.

Attached to this letter, you will find the marked-up version of the manuscript. The marked-up version highlights the changes that we have done regarding the first manuscript's version. Please, notice also that due to your comments below, the line, page, and figure numbers mentioned in our rebuttal letters to the referees have now slightly changed. To avoid confusion, we are also attaching an updated version of the rebuttal letters in which we are indicating the page and lines for each referees' comment in **green**.

Yours sincerely and on behalf of all co-authors,

Leandro Ponsoni

Editor Decision: Publish subject to minor revisions (review by editor) (15 Apr 2020) by Petra Heil

Comments to the Author:

Dear Dr Ponsoni.

Thank you for ypur submission tc-2019-257 to TC.

"Statistical predictability of the Arctic sea ice volume anomaly: identifying predictors and optimal sampling locations"

Pls review the comments of both reviewers carefully and address these together with the ones outlined here in your revised version.

Editor's general comments:

* End of Abstract: Add info on how much of the SIV var anomaly is achieved by six well-placed data locations.
In the new manuscript's version, the abstract is slightly reformulated and it contains this information.

* 1-24: Add what is impacted: "to bring significant impacts" --> To what? Then connect to the follow-on effects (your "Regionally, native"). --> You might need to bring your "global" impacts forward and then discuss the "regional" ones, as the former are related to the climate system, the others are follow
We have improved the first paragraph of the manuscript. Among other improvements, we are now bringing more "impacts" to the discussion. As you suggested, we have first addressed the global impacts and so the regional ones. That was indeed a good point. Thank you for pointing it out.

* 2-10: "meltdown" is very strong and inappropriate wording. Suggest to change, i.e., "intense sea-ice loss" or "rapid sea-ice loss".
We have replaced “meltdown” by “intense sea ice loss”.

* 2-19: Swart et al., [2015] are really about the relevance of internal variability in coupled climate (CMIP style) models. The statement "it has been already shown that trends in the pan-Arctic sea ice extent can be masked by its long-term variability (Swart et al., 2015)" is not correct. Pls rephrase.
We have revisited Swart et al. [2015] and we agree with this comment. We have reformulated the paragraph. This sentence was dropped from the text since it does not bring impact to the paragraph’s content.

* There are some resolved issues on the underperformance of OHT, especially in conjunction with SST as predictant. Would you kindly expand in the "Discussion" section?
In the new version of the manuscript, there is a full paragraph dedicated to this point in Sec. 4 (Discussion).

* The early part of your paper promises an analysis on how spatial model resolution affects the SIV prdictablity, but the manuscript do not follow through on this. Pls explore this and discuss in the "Discussion" section.
In the introduction of the manuscript, we have raised the question of whether or not the results are model dependent, in particular, whether they are sensitive to horizontal resolution. We have shown that indeed model resolution has an impact on the results. Depending on the analyzed metrics, the resolution impacts positively or negatively the statistical prediction. However, at this stage, we do not have a clear understanding of this point, and further investigation is needed to better understand the impact of model resolution on the SIV statistical predictability. We have highlighted all these aspects in the discussions section (Sec. 4).

* Pls separate the "Discussion" section from "Conclusion" section.
Done.

Minor comments:

1-2: Change "6" and "3" to "six" and "three".
Done.

1-3: Change "2" to "two".
Done.

1-6: Italize "in situ". -- Throught manuscript.
Done for all instances in the text.

1-10: Change "4" to "four".
Done.

1-11: Change "enough" to "sufficient".
Done.

1-14: Change "6" to "six".
Done.

1-14: Change "As per 6 well-placed locations" to read "Adding further to six well place locations" and remove "by adding new sites".
Done.

1-15: Change "4" to "four"... Pls change all numericals (in text), that are less than twelve to "words".
Done. Also for the entire manuscript.

2-29: Remove "trivial".
Done.

2-31: Replace "for feeding" with "as input into".
Done, also in other instance throughout the text.

3-11: Replace "broadly" with "well"
Done.

3-20ff: Remove "Following this introduction ... observing sampling design."
Done.

3-30: Assuming that the horizontal resolution (and subsequently bathymetry, inputs, etc associated with the grid res) is the only difference, then:
a) Replace "model configurations" with "model horizontal grids", and
b) Remove "These configurations differ by their horizontal grid resolution (in both the atmosphere and ocean)." (line 3-31).
Done. Please, notice that this section was slightly reformulated to address the referees' comments.

4-3: Figure references in the body of a manuscript should not be a repeat of the figure caption. Instead figures should be referenced in the text to support a statement/description of the displayed parameter(s). --> Rephrase.
We agree with the editor. This issue is corrected in the new version of the manuscript.

4-24: Remove "best".
Done (I guess pg. 5, though).

4-25: Replaece "non–significant" with "insignificant".
Done (I guess pg. 5, though).

6-1: Regarding reference to "Table 1", pls refer to my comment at 4-3.
Solved issue.

6-Tab1: In caption remind the reader why there is no CorrCoef for OHT and SIV for the high-res models.
Done.

7-Fig2: The order of the sub-figures is illogical, pls relabel.
Done.

8-19: Change "at least not" to "especially not".
Please, notice that we have followed the recommendation of Referee #2 and we reformulated Sec. 2.4. Due to that, this change is no longer required.

8-30: Change "to feed the Eq. 4" to "in Eq. 4".
Done.

9-15: Change "we borrow the concept" to "we follow the concept".
Done.

9-25: Change "Figure 3 shows how would be the region of influence ... ensemble of datasets." to "The region of influence for a station arbitrarily placed at the North Pole, as defined by the ensemble of datasets, exhibits depatures from concentric reflecting the transpolar drift. (Fig. 3)."
This suggestion was incorporated into the text, but please notice that we have reformulated Sec. 2.4 as suggested by Referee #2.

14-4: Change "6 models" to "sixe model realisations".
Done.

16-Fig7: Why does the colour scheme for each subfig in the right hand column change? Please show for a single colour scheme.
The color scheme used in the right-hand column in Fig. 7 is coherent with the representation of the same regions of influence in Fig. 8. If the editor allows, we would like to keep in this way so that the comparison between Fig. 7 and Fig. 8 is straightforward.

22-29: Change "ALL" to "all".
Done. Also in other instances throughout the manuscript.

23-2: Change "by observing platforms" to read "by autonomous observing platforms".
Done.

23-4: Change "The first" to "The former".
Done.

23-4: Change "not turn out to" to "did not act as".
Done.

23-4: Change "predictor (at least 5 not when using monthly means)." to "predictor, at least not when using monthly means."
Done.

23-19: Change "that only 4 stations are enough to overpass" to "that as few as four stations are sufficient to pass".
Done.

24-7: Change "might be in a free-ice region in the future." to "might in the future be ice free."
Done.

24-11: Spell out "MOSAiC".
Done.

Statistical predictability of the Arctic sea ice volume anomaly: identifying predictors and optimal sampling locations

Leandro Ponsoni[1,2], François Massonnet[1,2], David Docquier[3], Guillian Van Achter[1], and Thierry Fichefet[1]

[1]Georges Lemaître Centre for Earth and Climate Research (TECLIM), Earth and Life Institute, Université catholique de Louvain, Louvain-la-Neuve, Belgium
[2]Fonds de la Recherche Scientifique – FNRS, Belgium
[3]Rossby Centre, Swedish Metereological and Hydrological Institute, Norrköping, Sweden

Correspondence to: Leandro Ponsoni (leandro.ponsoni@uclouvain.be)

Abstract. This work evaluates the statistical predictability of the Arctic sea ice volume (SIV) anomaly – here defined as the detrended and deseasonalized SIV – on the interannual time scale. To do so, we made use of ~~6~~ six datasets, from ~~3~~ three different atmosphere-ocean general circulation models, with ~~2~~ two different horizontal grid resolutions each. Based on these datasets, we have developed a statistical empirical model which in turn was used to test the performance of different predictor variables, as well as to identify optimal locations from where the SIV anomaly could be better reconstructed and/or predicted. We tested the hypothesis that an ideal sampling strategy characterized by only a few optimal sampling locations can provide ~~in situ~~ *in situ* data for statistically reproducing and/or predicting the SIV interannual variability. The results showed that, apart from the SIV itself, the sea ice thickness is the best predictor variable, although total sea ice area, sea ice concentration, sea surface temperature, and sea ice drift can also contribute to improving the prediction skill. The prediction skill can be enhanced further by combining several predictors into the statistical model. Feeding the statistical model with predictor data from ~~4~~ four well-placed locations is ~~enough~~ sufficient for reconstructing about 70% of the SIV anomaly variance. ~~An improved model horizontal resolution allows a better trained statistical model so that the reconstructed values approach better to the original SIV anomaly. On the other hand, if we look at the interannual variability, the predictors provided by numerical models with lower horizontal resolution perform better when reconstructing the original SIV variability. As per 6 well-placed locations, the statistical predictability does not substantially improve by adding new sites.~~ As suggested by the results, the ~~4~~ four first best locations are placed at the transition Chukchi Sea–Central Arctic–Beaufort Sea (158.0°W, 79.5°N), near the North Pole (40°E, 88.5°N), at the transition Central Arctic–Laptev Sea (107°E, 81.5°N), and offshore the Canadian Archipelago (109.0°W, 82.5°N), in this respective order. Adding further to six well placed locations, which explains about 80% of the SIV anomaly variance, the statistical predictability does not substantially improve taking into account that ten locations explain about 84% of that variance. An improved model horizontal resolution allows a better trained statistical model so that the reconstructed values approach better to the original SIV anomaly. On the other hand, if we look at the interannual variability, the predictors provided by numerical models with lower horizontal resolution perform better when reconstructing the original SIV variability. We believe that this study provides recommendations for the ongoing and upcoming observational initiatives, in terms of an Arctic optimal observing design, for studying and predicting not only the SIV values but also its interannual variability.

1 Introduction

The continuous melting of the Arctic sea ice observed in the last decades (e.g., **?????????**), associated with the respective reduction in total sea ice area (SIA) and volume (SIV), has ~~proven to bring~~ led to significant impacts at ~~regional and global scales. Regionally,~~ global and regional scales. Globally, the sea ice depletion is reported to impact some aspects of the weather at low- and mid-latitude regions, by means of both oceanographic (**??**) and atmospheric teleconnections (**??**), including the higher occurrence of extreme events (**????**). Regionally, high-trophic predators such as seabirds (**??**) and mammals (**?????**) are changing their foraging behavior and dietary preferences. At the same time, native communities have experienced a disturbance ~~of~~ in subsistence activities like fishing, crabbing and hunting (**?**)~~, which have also impacted other predators from seabirds (??) to mammals (?????)~~. Other pressing local issues are also bringing important implications for the Arctic countries such as the opening of new ~~sailing~~ ship routes (**?**), the development of the tourism industry (**?**) and the mineral resource extraction (**?**). ~~At global scale, the sea ice depletion is reported to impact some aspects of the weather at low- and mid-latitude regions, by means of both oceanographic (??) and atmospheric teleconnections (??), including the higher occurrence of extreme events (????).~~

Since this ~~meltdown~~ intense sea ice loss is projected to continue throughout the twenty-first century (e.g., **?**), the interest of the scientific community and policy makers on the sea ice variability and predictability is exponentially increasing, mainly in terms of SIV. The SIV is a primary sea ice diagnostic because it accounts for the total mass of sea ice. ~~In situmeasurements of sea ice thickness, that are needed for calculating the SIV , are far too expensive, being spatially~~ However, *in situ*- and/or satellite-based estimates of SIV are still sparse and temporally sporadic (**??**). ~~In addition, satellite observations present well-known limitations in the warmer seasons and, therefore, sea ice thickness is not made available year-round from the classical satellite campaigns, namely: ICESat (?) and CryoSat-2 (?).~~

~~Due to the lack of~~ Due to this lack of continuous long-term observations, the answer to the question of whether or not this ~~negative trend~~ decline in sea ice ~~affects~~ is affecting the interannual variability of the pan-Arctic SIV, and the other way around, is not clear yet~~, although it has been already shown that trends in the pan-Arctic sea ice extent can be masked by its long-term variability (?). Up to now, most of the long-term SIV variability and predictability studies rely on reanalyses and/or model outputs (e.g., ?). Even though~~. Although, recent model analyses suggest that this might be indeed the case (**?**). Despite the fact that atmosphere-ocean general circulation models (AOGCMs), including their sea ice component, are more and more complex nowadays, ~~in situ~~ being even used to estimate the quality of global observational datasets (**?**), *in situ* observations are still required for a more comprehensive model validation and also for assimilation purposes.

In order to respond to the need of having an improved observational system for better understanding the SIV variability, but at the same time minimize the costs required to do so, this work raises the hypothesis that "an ideal sampling strategy characterized by only few optimal sampling locations can provide ~~in situ~~ *in situ* data for statistically reproducing and/or predicting the SIV interannual variability". To test the hypothesis, this study follows three main ~~directions~~goals. First, we propose a statistical empirical model for predicting the SIV. Since we are mainly interested in predicting the interannual variability rather than the ~~trivial~~ seasonal cycle and the long-term trends, we will focus on the SIV without these two components – hereafter defined as

SIV anomaly. Second, we aim at inspecting the performance of a set of ocean- and ice-related predictor variables ~~for feeding~~ as input into the empirical model. Third, we intend to localize a reduced number of optimal sampling locations from where the predictor variables could be systematically sampled using oceanographic moorings and/or buoys. Sampling ~~in situ~~ *in situ* data at optimal locations or, in order words, by collecting data at locations in which most of the pan-Arctic SIV anomaly variability is captured by the predictor variables, makes it much more feasible to sustain a long-term programme of operational oceanography both from logistical and financial points of view.

To the ~~knowledge~~ best of the authors' knowledge, this study is the first to apply an empirical statistical model for supporting an optimal observing system of the pan-Arctic SIV anomaly, albeit a similar study was conducted by **?** a decade ago. However, these authors focused on the predictability of averaged Arctic sea ice thickness, based their results on a single model approach, as well as considered two predetermined sampling locations. Other previous works also applied statistical empirical models for predicting a range of Arctic sea ice properties (e.g., sea ice extent, area and concentration), for lead periods of up to one year, at regional and/or pan-Arctic scales (**?????????**). Unlike the statistical prediction of sea ice extent and area, which have longer and more reliable records of observations allowing the statistical models to be built on this data, the statistical prediction of SIV necessarily requires information from models. *In situ* measurements of sea ice thickness, that are needed for calculating the SIV, are far too expensive, while satellite observations present well-known limitations in the warmer seasons. Therefore, sea ice thickness is not made available year-round from the classical satellite campaigns, namely: ICESat (**?**), CryoSat-2 (**?**), and SMOS (**??**).

Thus, even though we claim that ~~in situ~~ *in situ* observations are crucial for understanding the SIV variability, our study makes use of outputs from ~~3~~ three AOGCMs. This is the only way to have continuous and ~~broadly~~ well distributed data of the predictand and some predictor variables, such as sea ice thickness. ~~Here we assume that the AOGCMs~~ The AOGCMs used in this work are cutting edge in terms of model physics and resolution (**?**) so that they fairly represent the thermodynamic and dynamic processes linking predictors to predictand~~, while the use of 3~~. The use of three different models attempts to assess the model dependence of our results.

To fully address the three overall goals described above, this study is guided by the following open questions: (i) What ~~are~~ is the performance of different pan-Arctic predictors for predicting pan-Arctic SIV anomalies? (ii) What are the best ~~in situ~~ *in situ* locations for sampling predictor variables to optimize the statistical predictability of SIV anomalies in terms of reproducibility and variability? (iii) How many optimal sites are needed for explaining a ~~large amount , that is to say, at least~~ substantial amount (e.g., 70% ~~(~~– an arbitrarily chosen threshold) of the original SIV anomaly variance? (iv) Are the results model dependent, in particular, are they sensitive to horizontal resolution? ~~Following this introduction (Section 1), the manuscript is organized as follows: Section 2 describes the AOGCMs, datasets and the methods (including the development of the statistical empirical model) used in our analyses. Section 3 presents the results which are further discussed in Section 5. This last section also highlights the main conclusions and draw some recommendations for an observing sampling design.~~

2 Data and methods

2.1 Model outputs

~~As argued above, this work can not be performed with actual observations and it follows therefore~~

This work follows a multi-model approach. It takes advantage of ~~6~~ six coupled historical runs from ~~3~~ three different AOGCMs, all conducted within the context of the High Resolution Model Intercomparison Project (HighResMIP; **?**). The HighResMIP is ~~inserted in the framework of~~ endorsed by the Coupled Model Intercomparison Project 6 (CMIP6; **?**) and its main goal is to systematically study the role of horizontal resolution in the simulation of the climate system. ~~In our study, we use 2 different model configurations for each of the 3 models. These configurations differ by their horizontal grid resolution (in both the atmosphere and ocean). The runs~~ These historical coupled experiments, referred to as “hist-1950” (**?**), start in the early 1950s ~~, spanning~~ and span for about 65 years until mid-2010s. ~~We extract monthly outputs from these model simulations~~They are not pegged to observed conditions and their initial state is achieved by control coupled experiments referred to as “control-1950”, also produced in the context of HighResMIP. “control-1950” runs are CMIP6-like pre-industrial control (“piControl”) experiments, but using fixed 1950s forcing (**?**) rather than 1850s forcing as in “piControl” (**?**). The forcing consists of greenhouse gases (GHG), including O_3 and aerosol loading provided by a 10-year mean climatology for the 1950s (**?**). A full description of the GHG concentrations used by CMIP6 and HighResMIP is presented in **?**.

In our study, we use two different model horizontal grids for each of the three models. Namely, the AOGCMs are: the version 1.1 of the Alfred Wegener Institute Climate Model (AWI-CM; **??**), the European Centre for Medium-Range Weather Forecasts Integrated Forecast System (ECMWF-IFS) cycle 43r1 (**?**), and the Global Coupled 3.1 configuration of the Hadley Centre Global Environmental Model 3 (HadGEM3-GC3.1; **?**). ~~Fig. ?? shows the absolute values and the anomalies (no long-term trend; no seasonal cycle) of the Arctic SIV time series from the 6 model outputs.~~

~~Sea ice volume time series from the 6 model outputs used in this work: (a) absolute values and (b) anomalies (no trend; no seasonal cycle).~~

A comprehensive comparison including these ~~3~~ three models and their respective specifications are presented by **?**. In short, AWI-CM is composed by the European Centre/Hamburg version 6.3 (ECHAM6.3) atmospheric model and by the version 1.4 of the Finite Element Sea ice-Ocean Model (FESOM; **??**). ECMWF-IFS is a hydrostatic, semi-Lagrangian, semi-implicit dynamical-core atmospheric model, while the ocean and ice components are composed by the version 3.4 of the Nucleus for European Modelling of the Ocean model (NEMO; **?**) and version ~~2~~ two of the Louvain-la-Neuve Sea-Ice Model (LIM2; **?**), respectively. Finally, HadGEM3-GC3.1 is built up with the same ocean model than ECMWF-IFS (NEMO; **?**), but version 3.6, the atmospheric Unified Model (UM; **?**) and the version 5.1 of the Los Alamos sea-ice model (CICE; **?**). Hereafter the models are simply referred to as AWI, ECMWF, and HadGEM3.

~~Overall, the 2~~ The two configurations from the same model keep the parameters identical, except for the resolution-dependent parameterizations (**?**). In terms of ocean–sea ice grid, both AWI versions (data source: ~~**??**~~**??**) use a mesh grid with varying resolution, in which dynamically active regions have finer resolution. The low-resolution version (AWI-LR) has a ~~resolution changing from 24 to 110 km , and~~ nominal resolution of 250 km (e.g., 129 km at 50°N and 70 km at 70°N), while the high-resolution version (AWI-HR) ~~changes from 10 to 60 km (?)~~has nominal resolution of 100 km (e.g., 67 km at 50°N and 36 km at 70°N). Nevertheless, both versions have a similar resolution of ∼25 km in the Arctic Ocean. For a better understanding of AWI's grid, the reader is referred to **?** (their Fig. 4a,b). Both ECMWF (data source: ~~**??**~~**??**) and HadGEM3 (data source: **??**) adopt the tripolar ORCA grid (**?**). The configurations with coarser resolution (ECMWF-LR and HadGEM3-LL) use ORCA1 with a resolution of 1°, while the versions with finer horizontal grid (ECMWF-HR and HadGEM3-MM) use ORCA025 with a resolution of 0.25°. In terms of time resolution, our results are all based on monthly outputs from these model simulations.

2.2 ~~Potential predictors~~

~~The next step for proceeding with the statistical predictions of the SIV anomalies is to identify potential predictor variables to be used in the empirical statistical model~~For the three models, the SIV time series from the versions with a coarser horizontal grid present higher mean values compared to their respective finer resolution versions (Fig. 1a). The differences between the two versions are about 4.52×10^3 km^3 and 2.56×10^3 km^3 for AWI and HadGEM3, but much larger to ECMWF (26.17×10^3 km^3). The standard deviations (STD) from the SIV anomalies indicate that interannual variabilities are also higher for the coarser grid versions (Fig. 1b). The difference between coarser and finer resolutions for AWI, ECMWF, and HadGEM3 are 0.30×10^3 km^3, 1.78×10^3 km^3, and 0.43×10^3 km^3. We recall that the term ~~"anomaly "~~anomaly in this work refers to the detrended and deseasonalized time series. In practical terms, the anomaly is calculated by excluding the individual trend ~~(~~provided by a second-order polynomial fit ~~)~~ of each individual month.

2.2 Potential predictors

In this section, ~~as a first assessment, we test the performance of different predictors by estimating their correlation against the predictand. This test is performed individually for each model output, which means to say that predictor variables from a certain model configuration are only used for predicting the SIV anomaly from this respective configuration~~we identify potential predictor variables for using as input into the empirical statistical model that predicts SIV anomalies. Apart from the condition that all predictor variables could be regularly sampled from observational platforms in the real-world, we only pre-selected variables which have the potential to impact the sea ice through dynamic and/or thermodynamic processes. Overall, two categories of predictors are tested: global variables, intrinsically represented by a single pan-Arctic time series, and local predictors, represented by ~~gridded data. Nevertheless, for this first assessment,~~ several gridded time series of the same variable. Here, predictor variables are also considered in terms of their anomaly.

In total, a set of seven predictors are considered for this preliminary inspection. three of them are global variables, that are: pan-Arctic SIV itself, pan-Arctic sea ice area (SIA) and Atlantic basin ocean heat transport (OHT) estimated at 60°N. The other four predictors are local variables organized in a gridded format, that are: sea ice thickness (SIT), sea ice concentration (SIC), sea surface temperature (SST) and sea ice drift (Drift). Fig. 2 shows an example case (AWI-LR) in which the predictand (SIV) is compared against the two intrinsic pan-Arctic predictors (Fig. 2a,b) and against the four gridded predictors (Fig. 2c–j). As a first test, we inspect the performance of pan-Arctic predictors by estimating their lag-0 correlation against the predictand. The

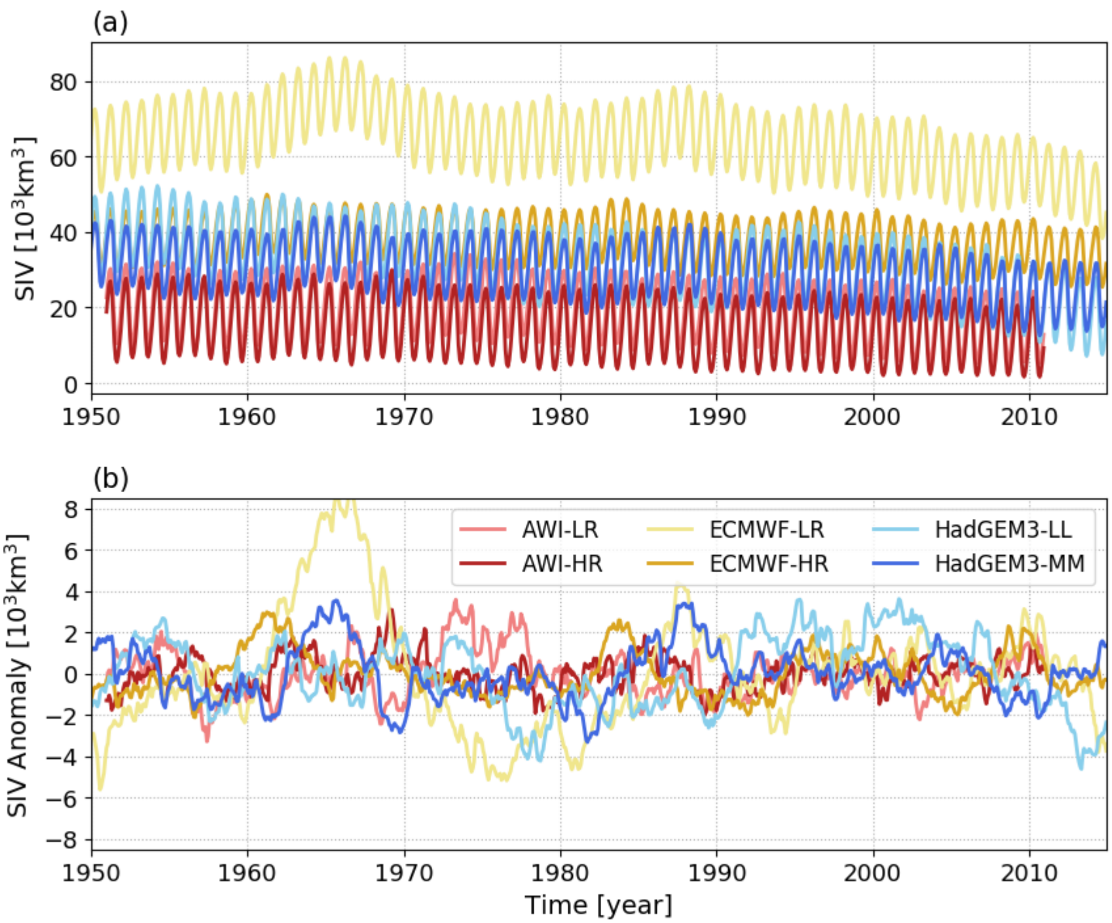

Figure 1. Sea ice volume time series from the six model outputs used in this work: (a) absolute values and (b) anomalies in which the long-term trends and the seasonal cycles were subtracted from the original time series.

correlation coefficients showed in the second (SIA) and third (OHT) columns of Table 1 indicate that SIA is a valid predictor for all model outputs, while OHT is significantly correlated only for the low-resolution versions of the models.

To obtain the same first assessment to the local predictors~~are considered as~~, the gridded values are reduced to their pan-Arctic ~~means~~average. To do so, the ~~gridded values~~ time series are twice normalized: first, by the grid area of each grid cell and, second, by the correlation maps with the predictand (**?**)~~. As suggested by **?** in~~, as shown in Fig. 2e,f,i,j. In the second normalization, the significant correlation coefficients from the different grid cells are used as normalizing factors (as it is the grid-cell area in the first normalization). The idea behind this second normalization is to take ~~the best~~ advantage of the correlations between predictand and predictors since the former is not necessarily correlated to the latter over the entire Arctic domain~~(Fig. ??). Notice that non–significant~~. Notice in the maps that insignificant correlation coefficients are set to zero (white regions) so that they do not ~~weight~~ weigh in the normalization ~~. Predictor variables are also used in terms of their anomaly (no trend; no seasonal cycle).~~

~~Apart from the condition that all predictor variables could be regularly sampled from observational platforms,we also considered only variables that have the potential to impact the sea ice through dynamic and/or thermodynamic processes. A set of 7 predictors are considered for this preliminary inspection. 3 of them are global variables,that are: pan-Arctic SIV itself,pan-Arctic SIA and Atlantic basin ocean heat transport (OHT)estimated at 60°N. The other 4 predictors are local variables provided by the AOGCMs in a gridded format and reduced to single time series as mentioned above, that are: sea ice thickness (SIT), sea ice concentration (SIC), sea surface temperature (SST) and sea ice drift (Drift). As an example, Fig. ?? compares the time series of predictand against pan-Arctic predictors (~~(Fig. 2e,f,i,j). The red lines in Fig. ~~??a,b,c,e~~2c,d,g,~~i), and also displays the respective correlation maps used for normalizing the regional predictors (Fig. ??d,f, h, j), for the AWI-LR output.~~ h show the respective SIT, SIC, SST and Drift anomalies reduced to their pan-Arctic averages, which are in turn significantly correlated with the predictand in all model outputs (Table 1~~shows the correlation coefficient estimated between predictand and predictors for all the models)~~.

Table 1. ~~Correlation~~ Lag-0 correlation coefficient estimated between the predictand (SIV anomaly) and a set of pan-Arctic potential predictors: SIA, OHT, SIT, SIC, SST, and Drift. The correlation coefficients between OHT and SIV anomaly for the high-resolution model versions are not shown since only statistically significant coefficients are displayed in the table. Regional predictors (SIT, SIC, SST and Drift) are represented by pan-Arctic averages. As for the predictand, all predictors are used with monthly time-resolution and in terms of their anomaly~~(no seasonal cycle; no long-term trend)~~.

Models	Predictors					
	SIA	OHT	SIT	SIC	SST	Drift
AWI-LR	0.64	-0.08	0.86	0.29	-0.57	0.15
AWI-HR	0.69	–	0.89	0.26	-0.50	0.31
ECMWF-LR	0.20	0.28	0.95	0.31	-0.12	-0.20
ECMWF-HR	0.24	–	0.63	0.37	-0.22	-0.13
HadGEM3-LL	0.63	-0.33	0.91	0.71	-0.54	-0.28
HadGEM3-MM	0.63	–	0.94	0.62	-0.45	-0.31

2.3 Statistical empirical models

The basis of our statistical empirical model (SEM) is a multiple linear regression model where the time series of the dependent variable (y) could be described as a function of the time series of the independent explanatory variables (x_i), as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon, \quad (1)$$

where β_0 is the constant y-intercept, β_k is the slope coefficients for each explanatory variable and ε is the error term (or residual) of the empirical model.

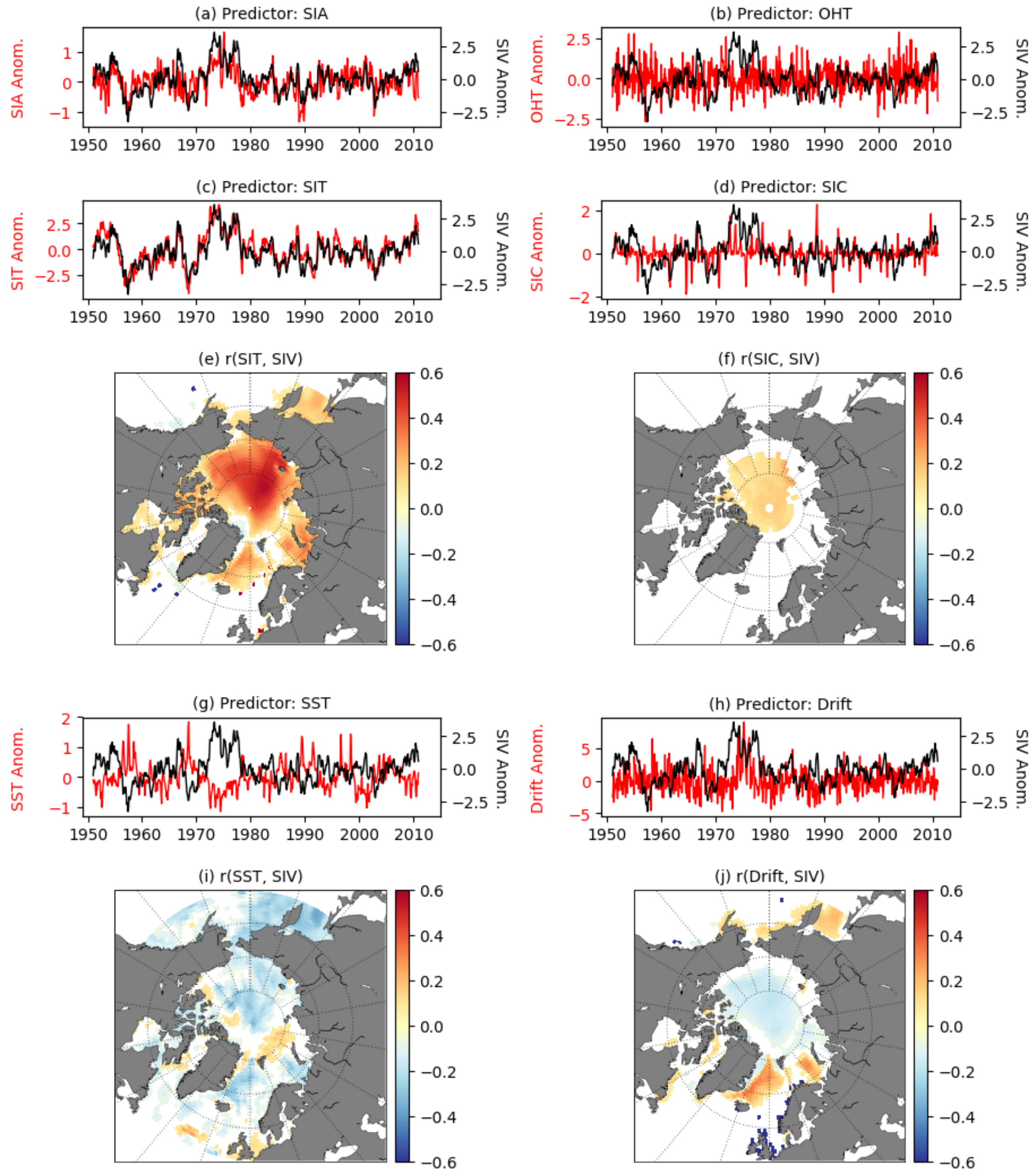

Figure 2. ~~Comparison~~ Lag-0 comparison between the time series from the predictand (SIV [10^3km^3]; black lines) and predictors: (a) SIA [10^6km^2], (b) OHT [PW], (c) SIT [m], (~~e~~d) SIC [%], (g) SST [°C] and (~~i~~h) Drift [km day^{-1}] (red lines). The correlation maps used for normalizing the regional predictors, as suggested by ?, are also shown: (~~d~~e) SIT, (f) SIC, (~~h~~i) SST and (j) Drift. Here, AWI-LR is merely used as an example case and not for a specific reason.

In our case, the reconstructed time series of SIV anomaly (~~SIV_{rec}~~SIV$_{rec}$) is based on the linear relationship between this variable and the predictors aforementioned in Section 2.2. If the SIV itself is also considered as a predictor, the multiple linear regression in Eq. 1 can be written as:

$$SIV_{rec} = \beta_0 + \beta_1 SIV + \beta_2 SIA + \beta_3 OHT + \beta_4 SIT + \beta_5 SIC + \beta_6 SST + \beta_7 Drift. \tag{2}$$

To bring robustness to the statistical reconstructions, the SEM is applied within a Monte-Carlo loop with 500 repetitions. In every repetition, 70% of the data are randomly selected for training (N_T) the SEM, while the remaining 30% are used for comparing (N_C) the original and the reconstructed SIV. In practical terms, ECMWF and HadGEM3 have 780 data points in time equivalent to the 780 months between Jan-1950 and Dec-2014 (720 for AWI; from Jan-1951 to Dec-2010) so that N_T = 546 monthly values are used for building the SEM and N_C = 234 values are used to evaluate how good is the SIV reconstruction (N_T = 504 and N_C = 216 for AWI). Since our main interest lies in the reconstruction of the SIV values, the metric used for comparing the original and reconstructed time series is the root mean squared error (RMSE). In this way, the score (Sc) of the reconstructed SIV can be represented by

$$Sc = \frac{1}{R}\sum_{r=1}^{R}\sqrt{\frac{\sum_{n=1}^{N_C}(SIV_{rec}(P) - SIV)^2}{N_C}}, \tag{3}$$

where R = 500 indicates the number of interactions in the Monte Carlo loop, P represents the (set of) employed predictor(s) and the index N_C emphasizes that only 30% of the data are used for comparison between original (SIV) and reconstructed SIV (SIV$_{rec}$) time series. An estimate of the Sc error (Sc$_{er}$) is given by the standard deviation calculated from the set of RMSEs given at every step of the Monte-Carlo scheme.

Two different approaches for applying the SEM are used in this work~~. In~~: First, in Section 3.1, we evaluate the individual and combined performances of the pan-Arctic predictors (intrinsic and averaged ones; see Section 2.2) for reconstructing the SIV anomaly at different months of the year (March and September), with a lag of ~~1~~ one to up to 12 months upfront. ~~In~~ Here, SIV itself is also allowed as an individual predictor to test the auto-prediction ability of this variable from lagged months. However, we are aware that SIV as a predictor could dominate the results since autocorrelation is expected to be stronger compared to the correlation with other variables. Therefore, SIV itself will not be used as a predictor in combination with other variables as generically described in Eq. 2 (see further Figs. 4h and 5h). Second, in Section 3.2, we make use of the SEM to support an optimal sampling strategy, but using the local predictors in their gridded format rather than their pan-Arctic averages, as the methodology described in Section 2.4. In this case, SIV itself is not used as a predictor at all.

2.4 Identifying optimal sampling locations

~~By identifying optimal sampling locations, we~~ We intend to spot a reduced number of sites from which predictor variables could offer an optimal representation of the pan-Arctic SIV anomaly. To ~~identify~~ identifying the 1st best location, a Score Map (Sc[i,j]) is ~~generated by calculating the Sc~~ created by applying the methodology described in Section 2.3 at each grid cell[i,j]~~,~~ ~~but now taking into account regional~~. However not all grid-point predictors (SIT[i,j], SIC[i,j], SST[i,j], Drift[i,j]) ~~rather~~

~~than~~ are necessarily used, but only the valid ones. That means, only predictors significantly correlated with the predictand are used. For instance, for the AWI-LR product, the SEM applied for a grid point placed off the eastern coast of Greenland will incorporate SIT, SST and Drift as predictors while SIC is disregarded, as suggested by the correlation maps plotted in Fig. 2e,f,i,j. SIA is the only intrinsic pan-Arctic ~~averages. From the predictors intrinsically represented by single time series (SIV,SIA,OHT),only SIA will be used because in the real world this variable is provided monthly from satellite measurements. SIV is disregarded for an obvious reason since this is the variable that we want to predict while having OHT from observations is a more complex task as it would require oceanic observations broadly distributed both in space and depth. Additionally,~~ predictor kept at this stage. The motivation for using SIA as a predictor is justified by the fact that this variable is already provided year-round by satellites so that it could be combined with *in situ* parameters in a real monitoring programme. OHT is not ~~a good predictor, at least not when it is used with monthly time-resolution~~used at this stage since it turned out that this predictor provides a relatively poor prediction to the predictand, as discussed further in Section 3.1. Also, from an observational point of view, sampling OHT is a very complex task that requires oceanographic observations well distributed both horizontally and in depth. SIV is disregarded for an obvious reason since this is the variable that we supposedly do not have and want to predict.

~~This method allows us to build Sc~~By following the approach above, the goal is to create a first Score Map (Sc[i,j]~~where the smaller the score, the better the representation of~~) from which the ~~pan-Arctic SIV anomaly. Hence, the most optimal location is here defined by the grid point where~~ 1st best location can be identified. In that Sc[~~i,~~i,j~~is minimum. In practical terms, the score maps will reveal clusters of grid points defining one region (or more) from where the SIV anomalies would be optimally reconstructed. After determining and fixing~~] the smaller the score, the better the grid point can reproduce the pan-Arctic SIV. The 1st ~~ideal location i_1,j_1, we can look for a 2nd i_2,j_2, a 3rd i_3,j_3, and so on~~ best location is the one represented by the smallest score in the Sc[~~i_k,j_k~~i,j]~~, best locations. However, every time that a location is identified, a region of influence surrounding this location is identifiedto avoid that different stations are placed nearby each other (see details below). In this approach, the regression described in Eq. 2, with k optimal locations, takes the following format:~~

~~$$SIV_{rec} = \beta_0 + \beta_2 SIA + \sum \beta_{P1[i_1,j_1]} P1[i_1,j_1] + \sum \beta_{P2[i_2,j_2]} P2[i_2,j_2] + \cdots + \sum \beta_{Pk[i_k,j_k]} Pk[i_k,j_k],$$~~

~~where the term $\beta_{Pk[i_k,j_k]} Pk[i_k,j_k]$ represents the product between the valid predictors $Pk[i_k,j_k]$, at the optimal location number k, and their respective slope coefficients $\beta_{Pk[i_k,j_k]}$. It is worthwhile mentioning that only valid predictors,which means only predictors significantly correlated with the predictand, are used to feed the Eq. 5. For instance, for the AWI-LR product, if a grid point placed off the eastern coast of Greenland is one of the N locations, the SEM incorporates SIT, SST and Drift as predictors while SIC is disregarded, as suggested by the correlation maps plotted in Fig. ??d,f,h,j.~~

~~For determining the 1st optimal location,this procedure is repeated independently for each of the 6 model outputs~~. Each of the six model outputs has its first Score Map. That means that each of the datasets provides its first optimal location (Sc[i_1,j_1]). ~~Not necessarily all the models will suggest the same locationas ideal. However, further~~In practical terms, the score maps will reveal not only a single best location, but clusters of grid points defining one (or more) region(s) from where the SIV anomalies would be optimally reconstructed (see further in Section 3.2.1~~, it will be shown that the different model outputs suggest relatively similar clusters of grid-points that can provide a skillful representation of the pan-Arctic SIV anomaly. Subsequently, aiming~~).

Aiming at spotting a single ~~first~~ 1st optimal location that better represents all datasets (ensemble 1st optimal location), we take the average of the ~~6~~ six score maps. To give the same weight for all datasets in the averaging, the individual score maps are scaled between ~~0 and 1~~ zero and one (ScNorm$_{[i,j]}$; **?**), as follows (Eq. 4):

$$ScNorm_{[i,j]} = \frac{Sc_{[i,j]} - Sc_{min}}{Sc_{max} - Sc_{min}}, \tag{4}$$

where the indexes min and max indicate the minimum and maximum values in the score map, respectively. Afterward, for having a coherent ~~gridded average, the 6~~ grid for averaging all normalized score maps, the six models are interpolated into a common ~~grid.~~ 1° × 1° grid. Besides the inherent different spatial grid-resolution of the models, this step has no impact on the results since the best-performing regions in the Score Maps are preserved (not shown). Finally, the 1st best ensemble sampling location is defined as the geographical coordinate where the mean ScNorm map presents its minimum value. ~~The advantage of this approach is to reduce~~ This approach has the advantage of reducing the model dependence of the results by relying on different datasets.

~~Notwithstanding, before departure for the identification of the 2nd ideal location~~ After determining and fixing the 1st ideal location [i_1,~~we borrow the concept of length scale (**??**)~~ j_1], we can look for a 2nd [i_2,j_2], a 3rd [i_3,j_3], and so on [i_k,j_k], best locations. However, every time that a new location is spotted, a region surrounding this point is also defined in order to avoid that ~~2~~ two optimal sites are placed near each other. To do so, we follow the concept of length scale (**??**). The length scale defines a radius where a certain gridded variable is well-correlated to the same variable from the neighboring grid points. In this work we do not use a radius, but a very similar approach: the correlation coefficient of our best local predictor at the selected location (SIT$_{\cancel{[i1,j1]}[ik,jk]}$; see Section 3.1) is calculated against the equivalent time series from all the other grid points (SIT$_{[i,j]}$). The region defined by the grid points with a correlation higher than $1/e$, a threshold for correlations below which the SIT is assumed to be uncorrelated to the point of interest, is used as a ~~buffer region—~~ restricting region. This region is hereafter defined as "region of influence". So, all ~~the~~ grid points enclosed ~~by~~ into the region of influence are automatically disregarded from being selected as ~~a 2nd location. Figure **??** shows how would be~~ the next optimal location. As an example, the region of influence for a station arbitrarily placed at the North Pole, as defined by the ensemble of datasets~~. Once the 2nd location is identified for all datasets, we repeat the procedure described above for determining a single 2nd optimal location . This iterative approach is also followed for the identification of the 3rd optimal site, and so on.~~ , exhibits departures from concentric reflecting the transpolar drift (Fig. 3).

In this approach, the regression described in Eq. 2, with k optimal locations, takes the following format:

$$SIV_{rec} = \beta_0 + \beta_2 SIA + \sum \beta_{P1[i_1,j_1]} P1[i_1,j_1] + \sum \beta_{P2[i_2,j_2]} P2[i_2,j_2] + \cdots + \sum \beta_{Pk[i_k,j_k]} Pk[i_k,j_k], \tag{5}$$

where the term $\beta_{Pk[i_k,j_k]} Pk[i_k,j_k]$ represents the product between the valid predictors $Pk[i_k,j_k]$, at the optimal location number k, and their respective slope coefficients $\beta_{Pk[i_k,j_k]}$. It is worthwhile mentioning that only valid predictors, which

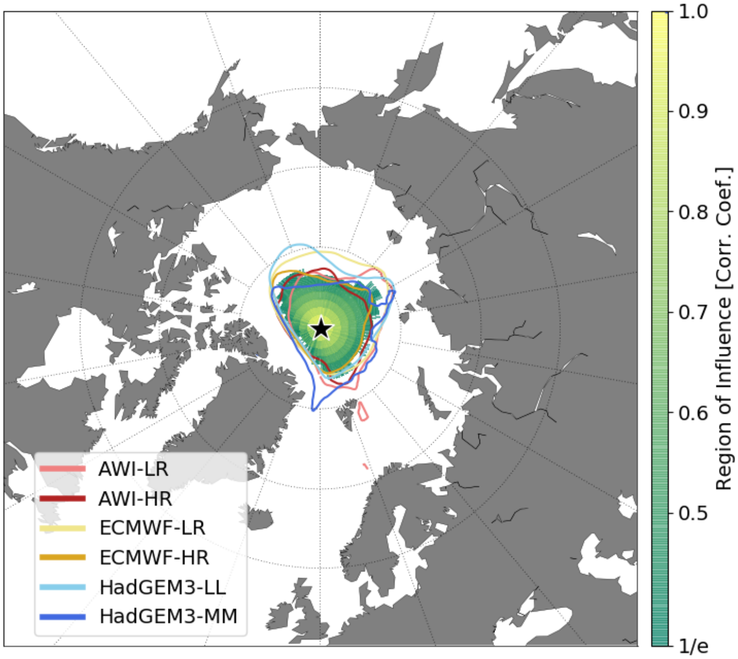

Figure 3. Region of influence for a station arbitrarily placed at the North Pole (black star) as defined by each model (colorful lines) and by the averaged region of influence from the different models (shades of green to yellow).

means only predictors from grid points placed outside the region of influence defined by previously selected points, and that are validated by the correlation map criterion, are used in Eq. 5.

3 Results

3.1 Statistical predictability of SIV anomaly: pan-Arctic predictors

In this section, the statistical predictability of the SIV anomaly is quantitatively evaluated by considering leading periods of ~~1~~ one to 12 months upfront. Also, the predictive performance of ~~7~~ seven pan-Arctic predictors is tested. The predictors are SIV itself, SIA, OHT, SIT, SIC, SST and Drift. Here, we focus on the months with relatively large (March; Section 3.1.1) and reduced (September; Section 3.1.2) SIV at the end of the winter and summer, respectively.

3.1.1 Statistical predictability of March SIV anomaly: pan-Arctic predictors

Figure ~~??~~ 4 displays the predictive performance (quantified by the RMSE) of different predictors for estimating March SIV anomalies. The SIV itself is the best predictor variable and its score gradually increases from 12 (Sc = 1.0 $\times 10^3$km^3) to ~~4~~ four (Sc = 0.68 $\times 10^3$km^3) leading months. During this period the mean performance for the ensemble of models increases by about

32%. As per ~~3~~ three leading months, from December to February, the predictive capacity substantially improves by 43% (Sc = 0.57 $\times 10^3$km^3), 59% (Sc = 0.41 $\times 10^3$km^3) and 77% (Sc = 0.23 $\times 10^3$km^3), respectively (Fig. ~~??~~4a).

The second best predictor is the SIT, which has performance similar to the SIV predictor from about 12 to ~~9~~ nine leading months (ensemble mean Sc = 1.02 $\times 10^3$km^3, 1.03 $\times 10^3$km^3, 1.0 $\times 10^3$km^3; Fig. ~~??~~4d). Nevertheless, its score remains relatively stable and improves only by about 25%, from May to February (Sc = 1.0 and 0.75 $\times 10^3$km^3). SIC (Fig. ~~??~~4e), SST (Fig. ~~??~~4f) and Drift (Fig. ~~??~~4g) have poorer performance compared to SIT, but similar behavior with the score slightly improving over time until ~~1~~ one leading month.

SIA (Fig. ~~??~~4b) is a valid predictor for AWI and HadGEM3 models, but it does not seem to be the case for ECMWF versions. Finally, OHT showed to be a poor predictor in terms of monthly predictability. For most of the leading months and models, the statistical reconstruction is not significant when provided by this predictor (Fig. ~~??~~4c).

A way of improving further the statistical predictability is to use several predictors at once. Figure ~~??~~4h shows the case where all the aforementioned predictors (except SIV) are used by the empirical model. For this configuration, the predictive skill is still 10% lower than the case where SIV is standing alone as a predictor, but it is about 10% better than the reconstructions provided only by the SIT.

The inter-model comparison does not show a conclusive answer to the question of whether or not the model resolution plays a role in the statistical predictability of March SIV anomalies. Overall, AWI-HR predictors are more skilled than AWI-LR predictors, though the opposite is observed for HadGEM3. For the ECMWF versions, the SIV anomalies from EMCWF-HR present better reproducibility, while ECMWF-LR presents much larger errors. Note that ECMWF-LR has a mean state characterized by a much thicker sea ice and, consequently, higher variance (see Fig. ~~??~~1). This is the reason that makes ECMWF-LR an outlier compared to the other ~~5~~ five model outputs for this and other results found in this manuscript (see further discussion in Section ~~5~~4).

3.1.2 Statistical predictability of September SIV anomaly: pan-Arctic predictors

A similar scenario compared to March is found for the September SIV anomaly predictability (Fig. ~~??~~5). The best predictor is the SIV itself (Fig. ~~??~~5a) for which the predictive skill improves by about 83.6% from June to August (Sc = 1.16 $\times 10^3$km^3 and 0.19 $\times 10^3$km^3). This improvement is mainly attributed to the ~~3~~ three months before September: Sc = 0.71 $\times 10^3$km^3, 0.44 $\times 10^3$km^3 and 0.19 $\times 10^3$km^3 for June, July and August, respectively. The second best predictor is SIT (Fig. ~~??~~5d), while SIC (Fig. ~~??~~5e), SST (Fig. ~~??~~5f) and Drift (Fig. ~~??~~5g) present an intermediate performance. For the former ~~4~~ four predictors, the ensemble mean Sc slightly improves from 12 to ~~1~~ one leading months in about: 28.8% (Sc = 1.04 $\times 10^3$km^3 and 0.74 $\times 10^3$km^3), 15% (Sc = 1.40 $\times 10^3$km^3 and 1.19 $\times 10^3$km^3), 29% (Sc = 1.26 $\times 10^3$km^3 and 0.90 $\times 10^3$km^3) and 24% (Sc = 1.46 $\times 10^3$km^3 and 1.11 $\times 10^3$km^3), respectively. Not all tested predictors are statistically significant for reproducing the SIV anomalies. Again, this is the case for OHT (Fig. ~~??~~5c). SIA also presents poor performance for some models and leading months (Fig. ~~??~~5b). Another resemblance to March predictability is the relatively poor performance presented by the predictor variables from ECMWF-LR.

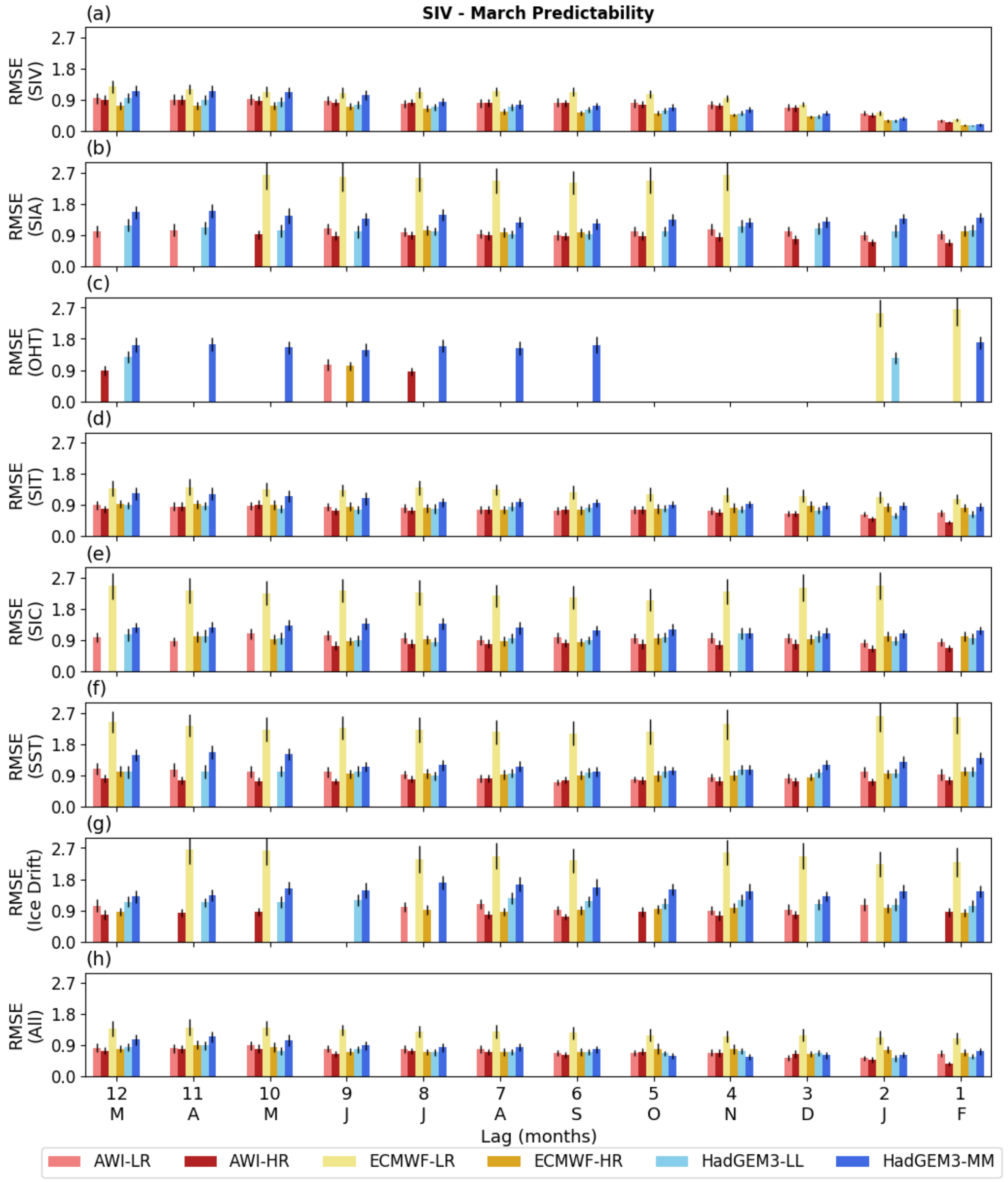

Figure 4. Statistical predictability of the March SIV anomalies, estimated from 12-leading months and quantified by the RMSE (10^3km^3) calculated between the original and reconstructed time series (Sc), as prescribed by ~~7~~ seven predictors: (a) SIV itself, (b) SIA, (c) OHT, (d) SIT, (e) SIC, (f) SST, (g) Drift. The predictions employing ~~ALL~~ all predictor variables (except the SIV itself) are displayed in (h). The vertical black lines indicate the error as provided by the 500 Monte Carlo simulations. The statistical predictability follows the methodology introduced in Section 2. Missing vertical bars mean that the statistical reconstruction is not statistically significant. The long-term trend and seasonal cycle are excluded from both predictand and predictors.

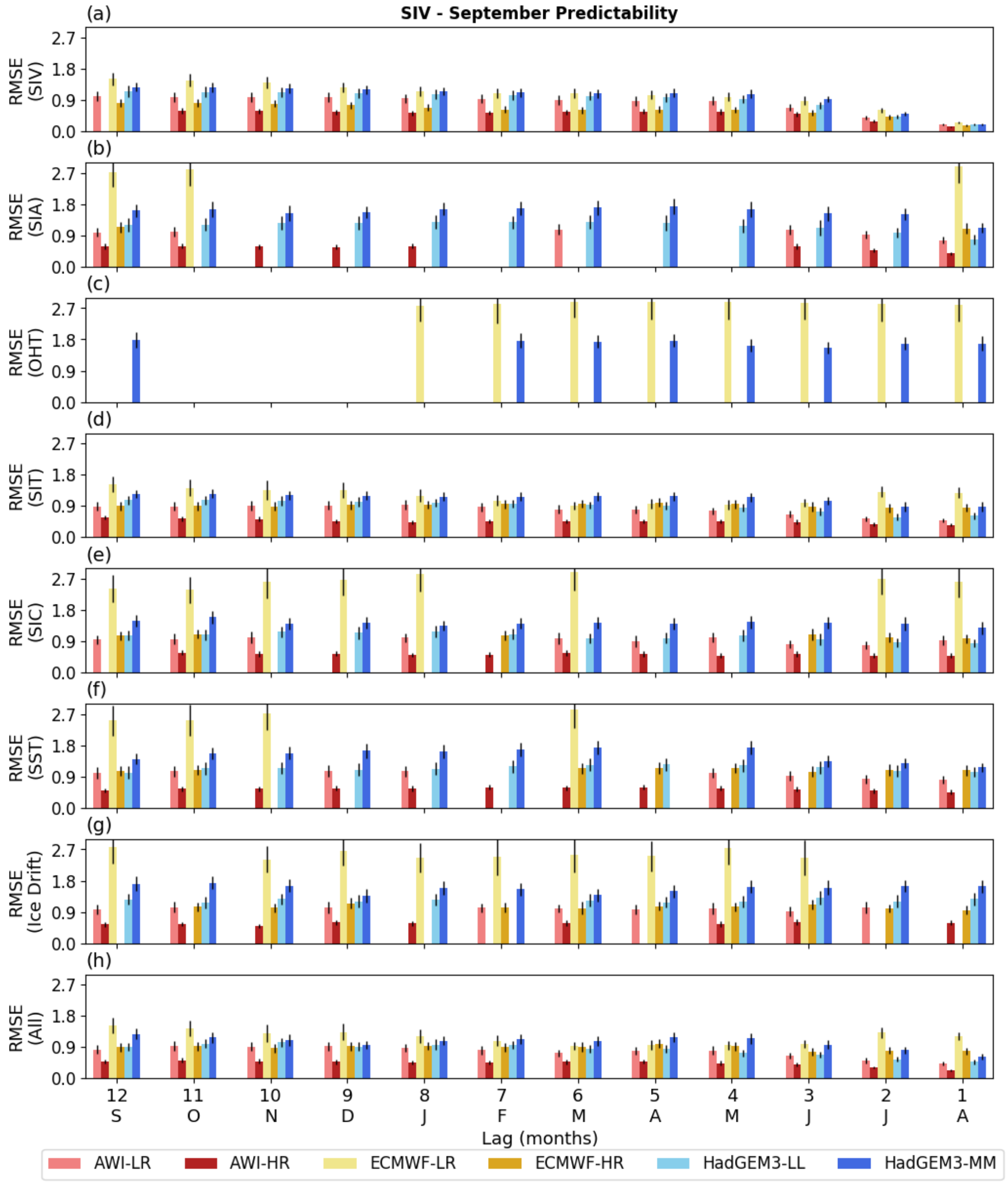

Figure 5. Statistical predictability of the September SIV anomalies, estimated from 12-leading months and quantified by the RMSE (10^3km^3) calculated between the original and reconstructed time series (Sc), as prescribed by ~~7~~ seven predictors: (a) SIV itself, (b) SIA, (c) OHT, (d) SIT, (e) SIC, (f) SST, (g) Drift. The predictions employing ~~ALL~~ all predictor variables (except the SIV itself) are displayed in (h). The vertical black lines indicate the error as provided by the 500 Monte Carlo simulations. The statistical predictability follows the methodology introduced in Section 2. Missing vertical bars mean that the statistical reconstruction is not statistically significant. The long-term trend and seasonal cycle are excluded from both predictand and predictors.

3.2 Statistical predictability of SIV anomaly: regional predictors

In this section, the empirical statistical model is used for supporting an optimal sampling strategy by following the methodology described in Section 2.4. To do so, we combine the local predictors at every grid-point rather than use their pan-Arctic averages. The reasoning behind this approach lies in the hypothesis that the statistical empirical model can fairly reproduce and/or predict the SIV anomalies if a few optimal locations provide ~~in situ~~ *in situ* measurements from the predictor variables. These ~~in situ~~ *in situ* observations can be applied concomitantly with predictors that are continuously measured by satellites as the pan-Arctic SIA and the local SIC.

Here we assume that numerical models are able to reproduce the main physical processes behind the interactions among predictand and predictors. Practically, we will take into account ~~4~~ four local predictors that are SIT, SIC, SST and Drift, and ~~1~~ one pan-Arctic predictor that is SIA, although it is worthwhile reminding that only predictors significantly correlated with the predictand will be incorporated to the statistical model. As per the results of Section 3.1, the OHT will not be included as predictor variable due to its poor capacity to provide a skillful prediction, being reinforced by the difficulties associated with the ~~in situ~~ *in situ* sampling and estimation of this variable.

3.2.1 Optimal sampling locations

For each of the ~~6 models~~six model realizations, score maps (Sc[i,j]; Eq. 3) were determined with the aim of spotting the location that can better reproduce the SIV anomalies as shown in Fig. ~~??~~6. This location is so defined as the grid point with minimum RMSE calculated between the original and reconstructed time series (Sc[$i1$,$j1$]; black stars in Fig. ~~??~~6). The spotted ideal location for AWI-LR, AWI-HR, and HadGEM-LL (Fig. ~~??~~6a,b,e) are relatively close to each other, separated by a maximum of ∼600 km. Even though ECMWF-LR, ECMWF-HR, and HadGEM3-MM (Fig. ~~??~~6c,d,f) suggest optimal locations that are placed farther from the sites suggested by the other datasets, their score maps still suggest a relatively good skill (low RMSE values) at the common region occupied by the ~~3~~ three previous referred models. This fact justifies further the multi-model approach used in this work.

The RMSEs (and associated STD from the Monte-Carlo scheme) calculated between the original SIV anomalies and the SIV anomalies reconstructed by the ~~ESM~~SEM, feed with predictor variables from the 1st optimal location (black stars in Fig. ~~??~~6), are shown in the mid column of Table 2. Based on those values, predictor variables from the AWI systems can better reproduce the SIV anomalies compared to the predictors from HadGEM and ECMWF. For the ~~3~~ three models, the high-resolution version provides better statistical predictability.

A common score map, with the indication of a common 1st optimal location placed at the transition Chukchi Sea – Central Arctic – Beaufort Sea (158.0°W, 79.5°N), is shown in Fig. ~~??~~7a. This common location is found through the ensemble mean of the scaled individual score maps, following the methodology described in ~~Sec.~~Section 2.4. If we now come back to the score maps in Fig. ~~??~~ 6 and retrieve the RMSE from that common location in Fig. ~~??~~7a, we find the values displayed in the right column of Table 2. The predictive skill drops by about 10% when the common point is chosen for all models, except for AWI-LR which presents similar results for the two locations. Those values also reinforce that, at least for this 1st location, the

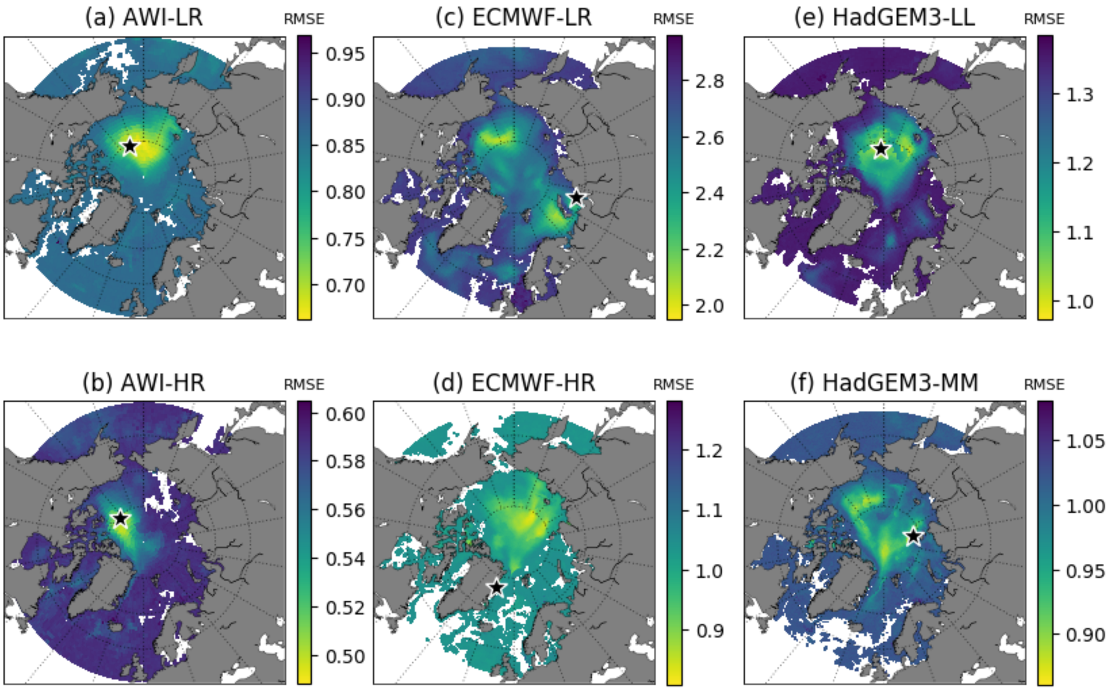

Figure 6. Score Maps (Sc[i,j]) represented by the RMSE (10^3km^3) calculated at every grid cell between the original and the reconstructed SIV anomalies. The smaller the RMSE error (shades of yellow), the higher the performance of the grid point for reconstructing the SIV anomaly. The black star indicates the 1st optimal location for each model Sc[$i1$,$j1$]. Notice that the colormap scale is different for each map.

predictors from the high-resolution outputs lead to a better predictive skill compared to the low-resolution predictors from their counterpart. Note that this was not the case when using pan-Arctic predictors in Section 3.1.

Once a 1st common optimal site is determined, we fix it for all datasets and so look for the 2nd best location. For that, the neighboring grid points which fell into the region of influence of the 1st best site are not considered as a second option. Fig. ~~??~~7b shows the 1st location's region of influence.

The procedure followed for identifying the 1st site is so repeated for the nth next locations. Aiming at improving the reconstruction of the SIV anomalies, every time that a new location is set, the valid predictors from this new point add to the predictors from the previous stations into the SEM. Fig. ~~??~~7c,e,g,i show the 2nd to the 5th optimal sites accompanied by their respective regions of influence (Fig. ~~??~~7d,f,h,j). The 2nd site is the one closest to the North Pole, from where it is separated by a distance of about 167 km. The 3rd, 4th and 5th points are placed at the offshore domain of the Laptev Sea in the transition with the Central Arctic, in the Central Arctic to the north of the Canadian Islands, and in the central domain of the Beaufort Sea, respectively.

If we think of an optimal observing framework, in which only a few observational platforms are deployed, Fig. ~~??~~ 8 represents an idealized scenario with the ~~10~~ ten best locations and their respective regions of influence. In such a context, the

Table 2. Mean RMSEs (and associated STDs) from the 500-Monte-Carlo realizations calculated between the original SIV anomalies and the SIV anomalies reconstructed by the ~~ESM~~SEM. We recall that in each Monte-Carlo realization 70% of the data is randomly used for training the SEM, while 30% is used for calculating the error. The first column shows the values for the case where the predictors are extracted from the individual optimal locations, while the second column shows the values found with predictors from the common optimal location.

Models	RMSE (Error) $\times 10^3$km^3	RMSE (STD) $\times 10^3$km^3
	1st Optimal Location	1st Optimal Location
	Individual location	Common location
AWI-LR	0.66 (±0.03)	0.67 (±0.03)
AWI-HR	0.49 (±0.02)	0.54 (±0.02)
ECMWF-LR	1.95 (±0.06)	2.11 (±0.09)
ECMWF-HR	0.81 (±0.03)	0.91 (±0.04)
HadGEM3-LL	0.97 (±0.04)	1.09 (±0.05)
HadGEM3-MM	0.86 (±0.05)	0.95 (±0.04)

selection of points respects the hierarchy of the regions of influence in a way that the 2nd site can not be placed within the region of influence #1 (shades of red), the 3rd point can not be placed within the regions of influence #1 and #2 (shades of red and purple), and so on. Note that with the proposed methodology, the regions of influence from the ~~10~~ ten first locations are covering almost the entire Arctic Ocean and adjacent seas, with exception of the Canadian Archipelago, the Kara Sea, and the Greenland Sea (see Fig. ~~??~~9). But even for the two later cases, the region of influence from other locations are partially covering these seas (Fig. 9; black line). The question of whether or not is indeed required all ~~10~~ ten locations to fairly predict the SIV anomalies, both in terms of anomaly values and variability, will be answered in the next sections.

Table 3 displays the geographical coordinates of the ~~10~~ ten locations as well as the Arctic sub-regions occupied by them, as identified in Fig. ~~??~~9. The division of the Arctic in sub-regions is based on the classical definition adopted by the broadly used Multisensor Analyzed Sea Ice Extent - Northern Hemisphere (MASIE-NH) product, which is made available by the National Snow & Ice Data Center (NSIDC). Most of the stations are placed within the Central Arctic (2nd, 4th, and 8th), or in the transition of this region with the Chukchi Sea (1st) and Laptev Sea (3rd), where the sea ice tends to be perennial. The 5th location is placed at the central part of the Beaufort Sea, the 6th and 9th stations are located at the offshore and inshore limits of the East Siberian Sea respectively, the 7th site is suggested to be at the Barents Sea off the Severny Island and, finally, the 10th station is occupying the near-coast side of the Laptev Sea.

3.2.2 Reconstructed SIV anomaly

Once the set of ideal locations are established, these sites are used to effectively reconstruct the entire time series of SIV anomalies from the ~~6~~ six model outputs, by taking into account only the valid predictors from each location. Again, we will make use of the RMSE to evaluate how good is our statistical prediction in terms of absolute values, but here we are also

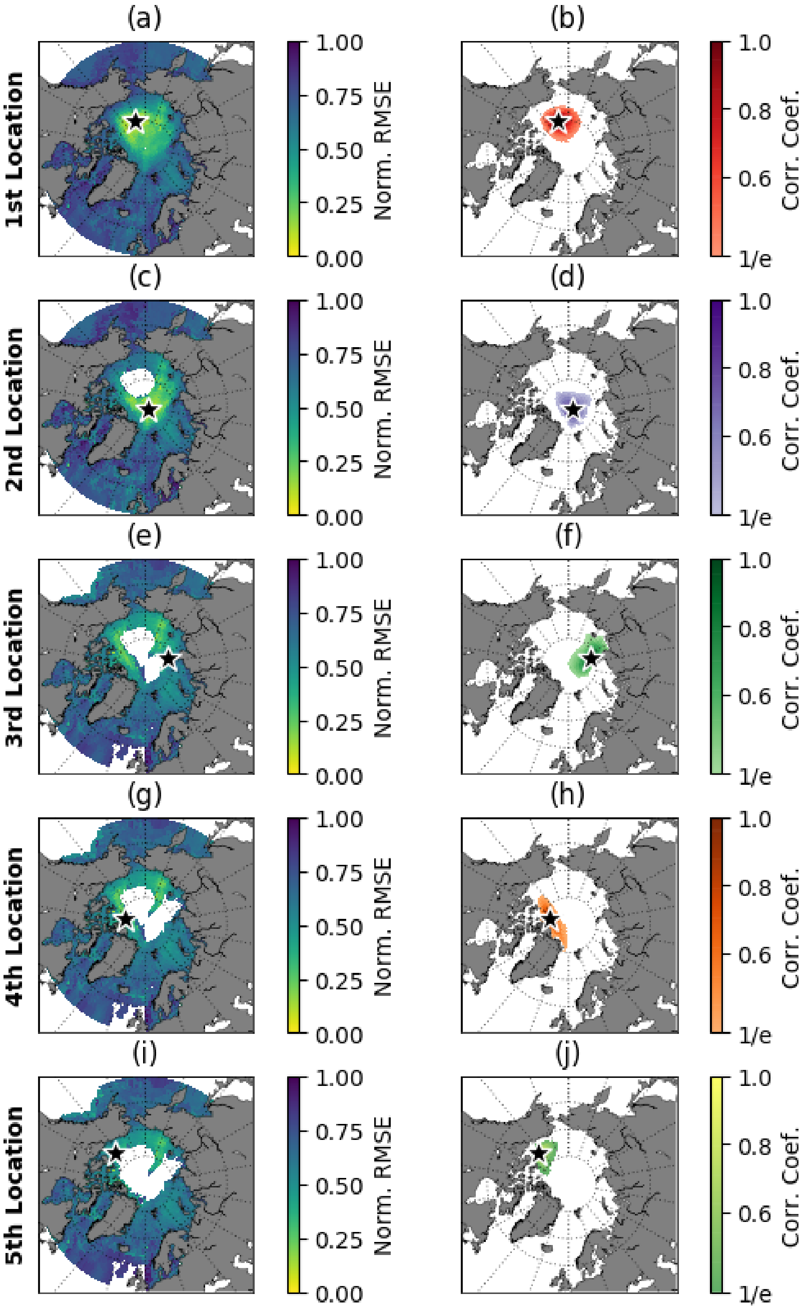

Figure 7. (a) Ensemble mean–normalized score map (ScNorm) for the 1st best sampling location. (b) The region of influence is defined for the 1st best location. The panels (c,d), (e,f), (g,h) and (i,j) represent the same as (a,b) but for the 2nd, 3rd, 4th and 5th best sampling locations, respectively.

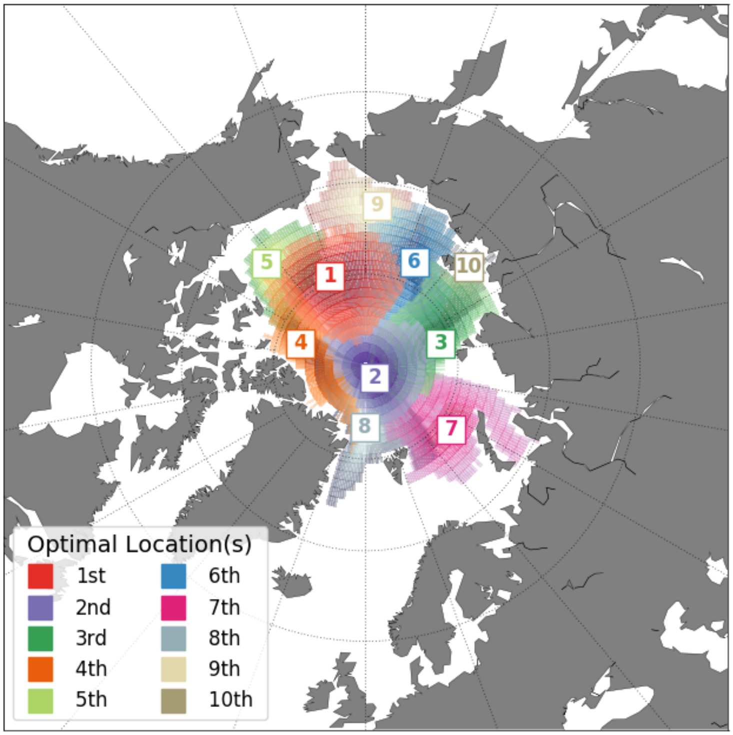

Figure 8. Optimal observing framework, as suggested by the ensemble of model outputs, for sampling predictor variables in order to statistically reconstruct and/or predict the Pan-Arctic SIV anomaly. The numbers indicate the 1st up to the 10th best observing locations in the respective order. The hatched area around each location (same color code) represents their respective region of influence. The selection of points respects the hierarchy of the regions of influence in a way that the 2nd point can not be placed within the region of influence #1 (shades of red), the 3rd point can not be placed within the regions of influence #1 and #2 (shades of red and purple), and so on.

interested in inspecting the ability of the empirical model to reproduce the full variability of the SIV anomalies. For that, apart from the RMSE, we also calculate the coefficient of determination (R^2) between the original and reconstructed time series.

Figure ~~?? compares~~ 10 provides a comparison at lag-0 between the original (black lines) and the reconstructed times series by taking into account the 1st (red lines), the ~~3~~ three first (green lines) and the ~~6~~ six first (blue lines) locations. For the first reconstruction, RMSE values are almost identical to the ones shown in the second column of Table 2 (see Fig. ~~??~~11a; y-axis=1). Again, for all ~~3~~ three models, the predictor variables from the higher resolution versions present better performance in reproducing the SIV anomaly values. The relatively poor skill of the ECMWF-LR predictors compared to the other ~~5~~ five systems is remarkable (~~Figure ??~~Fig. 10c).

Figure ~~??~~11a summarizes the RMSE values for the reconstructions conducted with data from ~~the~~ only the 1st up to all ~~10~~ ten combined locations. The pattern of better prediction skill for the models with higher grid resolution revealed by the 1st location remains when more sites are incorporated into the SEM. From the ensemble means the RMSE ($\times10^3$km^3) values are,

Table 3. Geographical coordinates for the first ~~10~~ ten optimal sampling locations (second and third columns). The fourth column informs the sub-regions in which each of the points are placed in (see Fig. ~~??~~9). The limits of the sub-regions are suggested by the National Snow & Ice Data Center (NSIDC).

Optimal Location	Latitude	Longitude	Sub-Region
#1	79.5°N	158.0°W	Chukchi Sea (CS)
#2	88.5°N	040.0°E	Central Arctic (CA)
#3	81.5°N	107.0°E	Central Arctic (CA)
#4	82.5°N	109.0°W	Central Arctic (CA)
#5	74.5°N	136.0°W	Beaufort Sea (BeS)
#6	77.5°N	155.0°E	East Siberian Sea (ESS)
#7	78.5°N	054.0°E	Barents Sea (BrS)
#8	83.5°N	001.0°W	Central Arctic (CA)
#9	72.5°N	176.0°E	East Siberian Sea (ESS)
#10	74.5°N	134.0°E	Laptev Sea (LS)

respectively, 1.06, 0.95, 0.90, 0.81, 0.78, 0.70, 0.65, 0.63, 0.60 and 0.59 for the reconstruction with ~~1 to 10~~ one to ten locations (black curve/points in Fig. ~~??~~11a). By excluding the outliers from ECMWF-LR, the previous RMSEs reduce to about 20% as shown by the gray curve-points in Fig. ~~??~~11a). For most of the datasets, the statistical reconstruction seems to improve better until the incorporation of the 5th to 6th locations, from when on the improvement seems to attenuate (~~Figure ??~~Fig. 11a).

Figure ~~??~~11b introduces a similar analysis but quantified by the R^2. Interestingly, for this metric, the ECMWF-LR is not outstanding from the others, and its predictors present a similar performance for reproducing the SIV anomaly variability. By account the reconstructions with ~~1 to 10~~ one to ten optimal sites, the ensemble means of R^2 values are: 0.53, 0.63, 0.67, 0.73, 0.75, 0.80, 0.81, 0.83, 0.84 and 0.84, respectively. These ensemble means suggest that the statistical empirical model could reproduce more than 60% of the SIV variability by using predictors from only the ~~3~~ three first optimal locations. AWI and HadGEM datasets indicate that ~~4~~ four locations are enough for reproducing more than 70% of the variability. With ~~6~~ six well-positioned sites, about 80% of the SIV anomaly could be explained as suggested by the ensemble mean (Fig. ~~??~~11b). As per the 6th station, the gain from adding new locations seems to be minimal (∼1%). Also interestingly is the fact that for the R^2 metric, the opposite from the RMSE is observed since the best performing predictors are the ones coming from the model's version with lower grid resolution.

In terms of used predictor variables, ~~Figure ??~~Fig. 11c reiterates that SIT is the most skillful of the local predictors. From the 60 cases that the SEM was applied (~~6 datasets, 10~~ six datasets, ten locations), SIT was used 59 times. SIT was not a valid predictor only for the 9th location in ECMWF-HR. SST and Drift were used in about two thirds (20 and 22 times, respectively), while SIC was used only in half (31 times) of the cases. If we look at the individual model outputs, HadGEM (the ~~2~~ two resolutions comprised) is the one in which the empirical model takes the best advantage of the available gridded

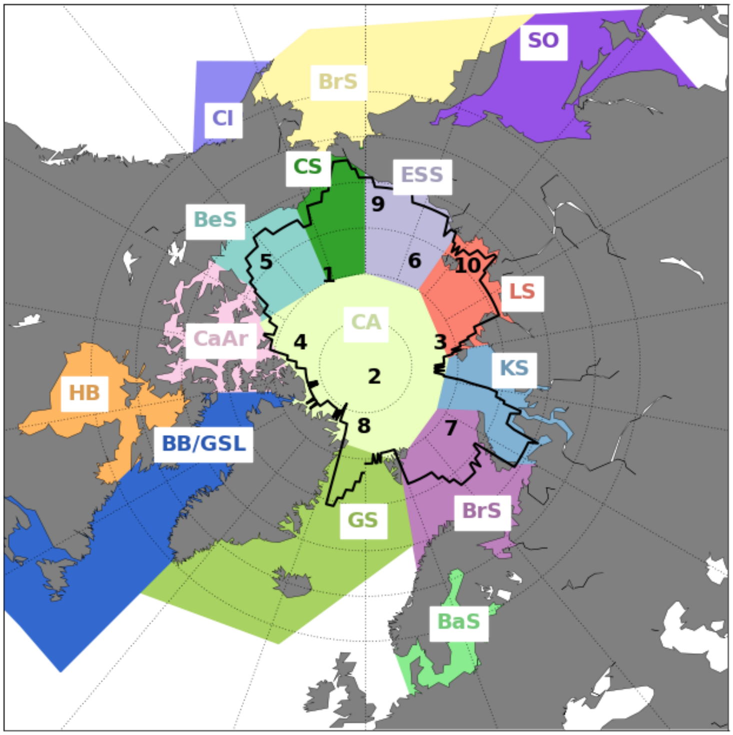

Figure 9. Optimal observing framework for sampling predictor variables in order to statistically reconstruct and/or predict the Pan-Arctic SIV anomaly. The numbers indicate the 1st up to the 10th optimal sites. Each of the colored areas represent an Arctic sub-region according to the Arctic subdivision suggested by the National Snow & Ice Data Center (NSIDC). The black line indicates the global region of influence defined in Fig. 8 (color-shaded areas). Acronyms: Beaufort Sea (BeS); Chukchi Sea (CS); East Siberian Sea (ESS); Laptev Sea (LS); Kara Sea (KS); Barents Sea (BrS); Greenland Sea (GS); Baffin Bay/Gulf of St. Lawrence (BeS); Canadian Archipelago (CaAr); Hudson Bay (HB); Central Arctic (CA); Bering Sea (BrS); Baltic Sea (BaS); Sea of Okhotsk (SO); Cook Inlet (CI).

predictors, having neglected one of them in only 15 out of 80 cases, while ECMWF and AWI have ignored predictors in 29 and 30 out of 80 cases, respectively.

To evaluate the performance and robustness of our SEM, the RMSE and R^2 calculated between the original and our-methodology-based reconstructed SIV anomalies (Fig. 11a,b) are compared against the same two metrics but now estimated by a simple multiple linear regression model having as input predictor data from randomly chosen locations (Fig. 12). For that purpose, 100 combinations of ten randomly chosen locations were determined. For each combination, the SIV anomaly is reconstructed with predictor data from the 1st location, the 1st–2nd, the 1st–3rd, ..., the 1st–10th locations. For the sake of fairness, we have used the same predictor variables from randomly locations placed only into the global region of influence represented by the black line in Fig. 9. The results show that the SIV reconstructions based on our methodology (and optimally selected locations)

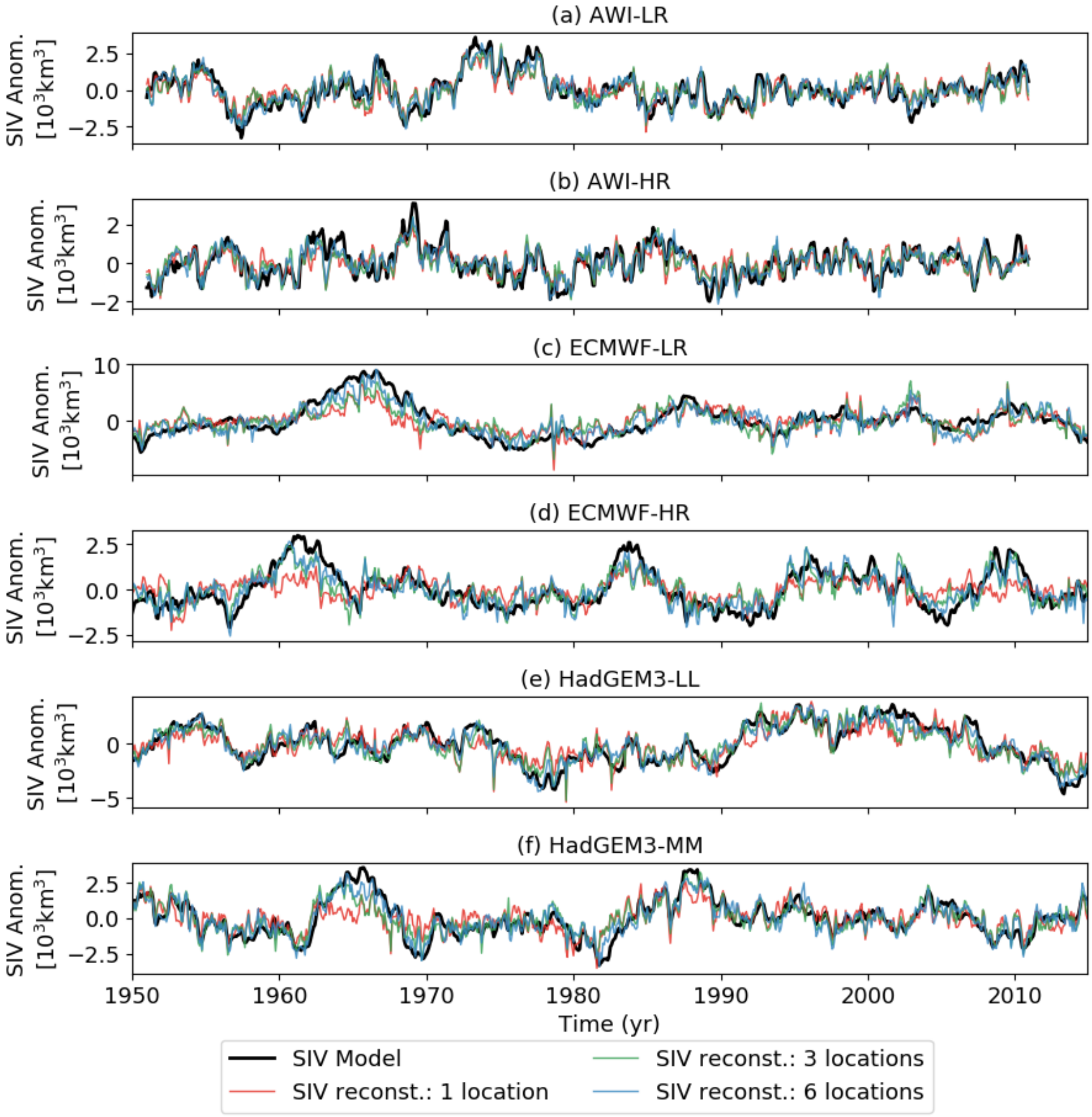

Figure 10. ~~Original~~ Lag-0 comparision between the original (black) and statistically reconstructed SIV anomalies. The reconstruction takes into account the 1st (red), the ~~3~~ three first (1st–3rd; green) and the ~~6~~ six first (1st–6th; blue) optimal locations: (a) AWI-LR, (b) AWI-HR, (c) ECMWF-LR, (d) ECMWF-HR, (e) HadGEM-LL and (f) HadGEM-MM. Notice the different scales in the y-axes.

are more skillful both in terms of RMSE and R^2. This is valid for all models, considering a single location or any combination of up to ten locations (Fig. 12).

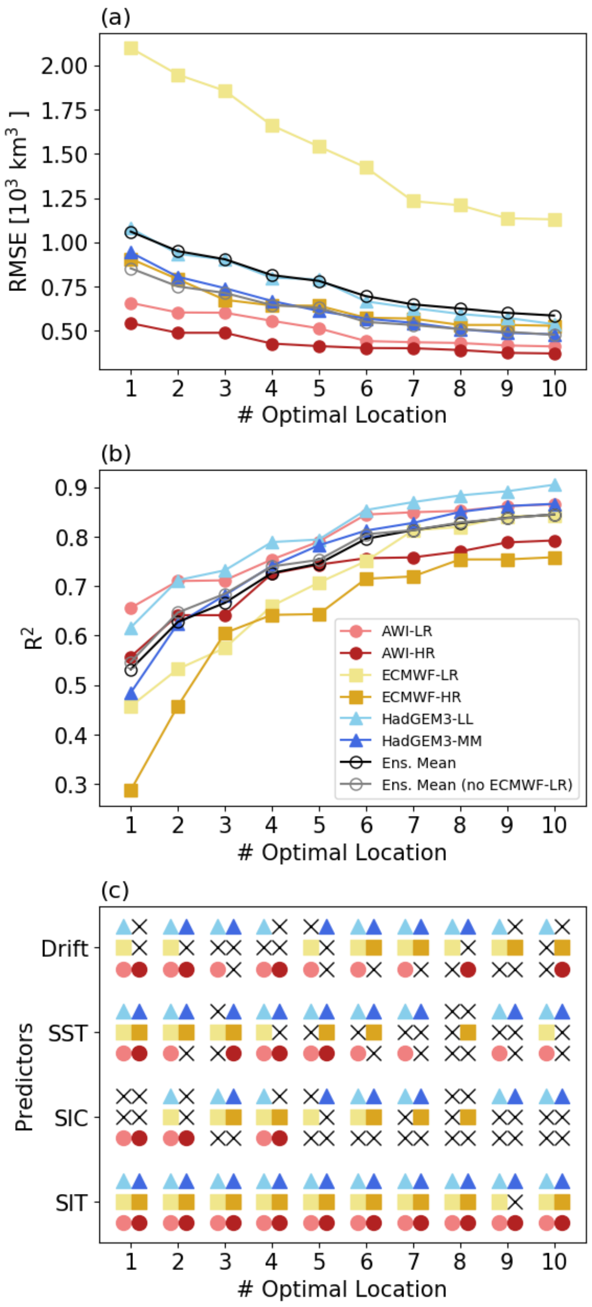

Figure 11. (a) RMSE (y-axis) estimated between the original and reconstructed time series by taking into account predictor variables from ~~1~~ one up to ~~10~~ ten optimally selected locations (x-axis). (b) Same as (a) but using R^2 (y-axis) to compare original and reconstructed time series. (c) Valid predictors, as determined by the correlation maps, retrieved from each targeted location. If a predictor is valid (y-axis), its respective symbol, as defined in the inset legend from (b), is plotted. A black cross indicates that the predictor is not valid at the respective location.

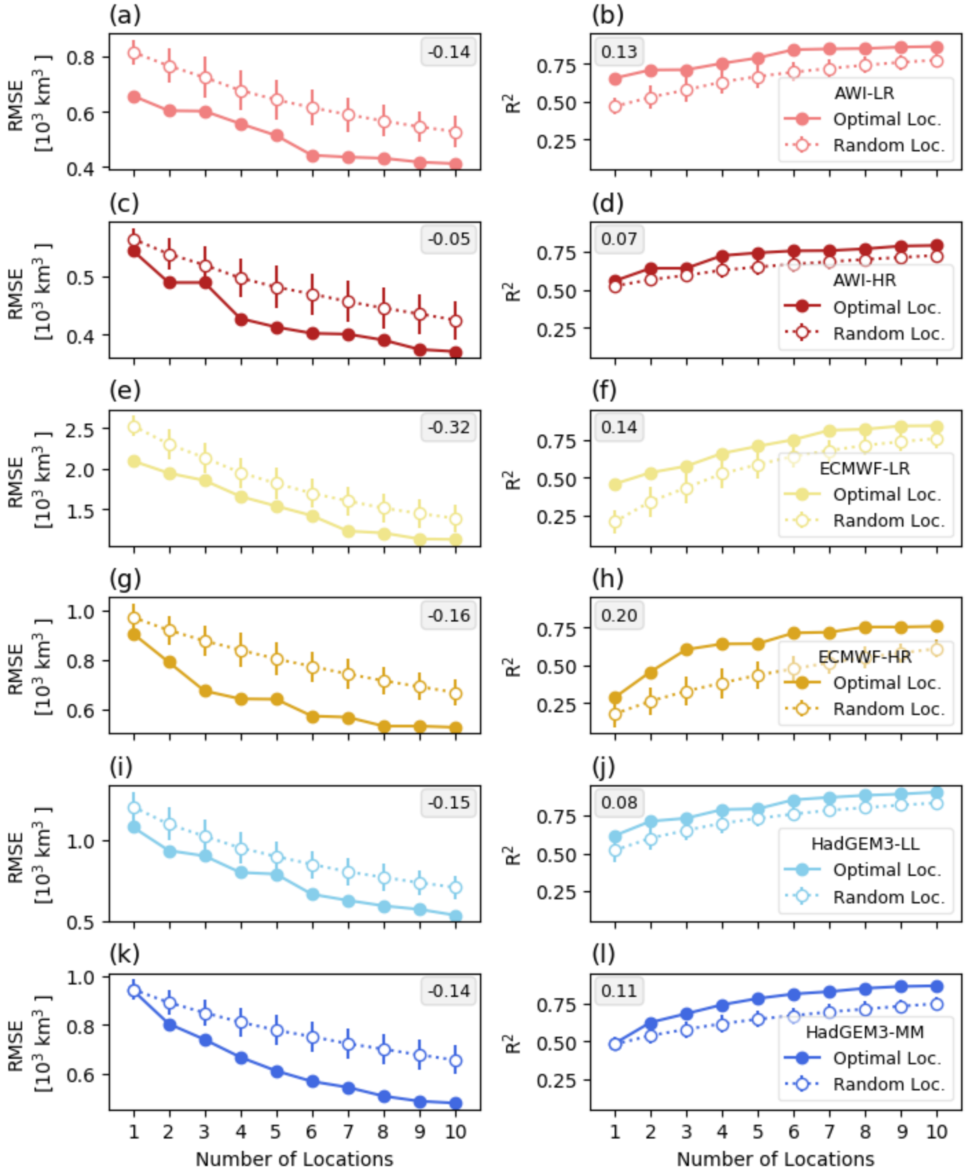

Figure 12. Root Mean Squared Error (RMSE; left column) and coefficient of determination (R^2; right column) calculated between the original and reconstructed SIV anomalies. The reconstructed SIV volume anomalies are based on the optimally selected locations following our methodology (full dots; same as in Fig. 11a,b), as well as by randomly chosen locations (empty dots). In the last case, 100 sets of ten randomly chosen locations are used. For each of the 100 sets, the SIV anomaly is reconstructed using data from the 1st, the 1st–2nd, the 1st–3rd, ..., the 1st–10th, locations. The random locations are all enclosed into the global region of influence defined (Fig. 9; black line). The vertical bars associated with the empty dots represent the one standard deviation from the 100 reconstructions. The inset numbers represent the average difference between the two curves shown in each sub-panel.

4 Discussion

In this work, we have introduced a statistical empirical model for predicting the Arctic SIV anomaly ~~(no trend; no seasonal cycle).~~ on the interannual time scale. The model was built and tested with data from ~~3~~ three AOGCMs (AWI-CM, ECMWF-IFS, and HadGEM3-GC3.1), each of which provided with ~~2~~ two horizontal resolutions, performing a total of ~~6~~ six datasets. We have first inspected the predictive skill of ~~7~~ seven different pan-Arctic predictors, namely: SIV, SIA, OHT, SIT, SIC, SST, and Drift. These predictors were tested since they have dynamical and/or thermodynamical influence on the SIV. The ~~3~~ three first are intrinsically represented by single time series, while the remaining are gridded variables that were reduced to mean pan-Arctic time series. From this first assessment, performed for the months of March and September, the results (Section 3.1) show that the best predictors are the SIV itself and the SIT, whilst SST, Drift, SIC and SIA provide some intermediate-skill predictions. In general, such results are valid for predictions performed from ~~1~~ one back to 12 leading months. For the SIV predictor, the skill substantially increases in the last ~~3~~ three leading months. For the remaining aforementioned predictors, the skill slightly improves from 12 to ~~1~~ one leading month. ~~OHT provided a~~

In contrast, OHT provided very poor predictive skill. **?** recently showed (their Fig. 12) a relatively good correlation between OHT and the SIV. However, these authors correlated annual averages of OHT against monthly values of SIV, but here we are considering monthly means for all predictors. Based on that, the results from both manuscripts suggest that the OHT has a cumulative impact on the sea ice throughout the year, which is not so remarkable when looking at individual months, even if several leading months are considered. One might wonder how SST is a relatively skillful predictor, while OHT not. We recall that the OHT tested as a predictor in this study is a remote parameter, which takes into account the seawater temperature (and meridional velocities) throughout the entire water column, calculated at 60°N for the Atlantic basin ocean (**?**). There are other potential candidates to explain why OHT is a poor predictor, as for instance model biases such as an overestimation of the stratification at the near-surface layer, which could attenuate the heat content being transported towards the Arctic Ocean. Nevertheless, this is a subject that requires a more detailed investigation.

From Section 3.1's results is also noticeable that the ECMWF-LR predictors present a relatively poor skill compared to the others. This is explained by the fact that this model has a mean state characterized by a much thicker sea ice (see Fig. ~~??~~1), impacting the ~~RMSE~~ RME used as a metric for evaluating the prediction skill.

That being said, we can recapitulate and objectively answer the first open question posed in the introduction of this manuscript:

(i) What ~~are~~ is the performance of different pan-Arctic predictors for predicting pan-Arctic SIV anomalies?

If we take into account the ensemble mean, and use the average RMSE calculated between original and reconstructed SIV time series (Section 3.1; Figs. ~~?? and ??~~4 and 5) for the last ~~3~~ three leading months as score, the best predictors for March are sorted in the following order: SIV (0.41×10^3km^3), SIT (0.78×10^3km^3), SIA (1.01×10^3km^3), SIC (1.10×10^3km^3), SST (1.15×10^3km^3), Drift (1.32×10^3km^3) and OHT (2.05×10^3km^3). The best predictors for September are sorted as: SIV (0.45×10^3km^3), SIT (0.76×10^3km^3), SST (0.96×10^3km^3), SIA (1.07×10^3km^3), SIC (1.12×10^3km^3), Drift (1.22×10^3km^3) and OHT (2.24×10^3 km^3). If ~~ALL~~ all predictors are used (except SIV itself), the averaged scores for ~~3~~ three leading months are 0.70×10^3km^3 for both March and September~~, respectively~~.

Once the statistical empirical model is developed and the potential predictor variables are identified, we made use of this information for recommending an optimal observing system. Such observations could eventually be performed in the framework of an operational oceanography program to continuously provide predictor data for the statistical model. So, we considered parameters that could be locally sampled by autonomous observing platforms (e.g., oceanographic moorings and/or buoys) as SIT, SST and Drift. It is fair also to consider the SIC and the pan-Arctic SIA since this information is regularly provided from satellite measurements. The OHT and the SIV are here disregarded as predictors. The ~~first did not turn out to be~~ former did not act as a skillful predictor~~(~~, at least not when using monthly means~~)~~. The second is the variable that we supposedly do not have and the one we want to predict. From a realistic point of view, our analyses were restricted to a maximum of ~~10~~ ten optimal locations, although a reduced number of stations would be already ~~enough~~ sufficient to fairly reproduce the SIV anomaly, and so to explain a large amount of its variance (see below). The results from Section 3.2 provide us with elements to answer the other ~~3~~ three open questions of this study, as follows:

(ii) What are the best *in situ* locations for sampling predictor variables to optimize the statistical predictability of SIV anomalies in terms of reproducibility and variability?

We have here identified ~~10~~ ten optimal locations. The exact coordinates of these locations are provided in Table 3 and also plotted in Figs. ~~?? and ??~~8 and 9. As suggested by the ensemble of model outputs, the 1st optimal location is placed at the transition Chukchi Sea–Central Arctic–Beaufort Sea (158.0°W, 79.5°N). The 2nd, 3rd and 4th best locations are placed near the North Pole (40°E, 88.5°N), at the transition Central Arctic–Laptev Sea (107°E, 81.5°N) and offshore the Canadian Archipelago (109.0°W, 82.5°N).

(iii) How many optimal sites are needed for explaining a ~~large~~ substantial amount (e.g., ~~that is to say, at least~~ 70% – an arbitrarily chosen threshold) of the original SIV anomaly variance?

By considering an arbitrary threshold of 70%, the systems AWI-LR (75%), AWI-HR (73%), HadGEM3-LL (79%) and HadGEM3-MM (74%) suggest that ~~only 4 stations are enough to overpass~~ as few as four stations are sufficient to pass this threshold, what is also confirmed by the ensemble mean (73%). Even though the ECMWF predictors have slightly low skill, they are still not far from the threshold: ECMWF-LR (66%) and ECMWF-HR (64%). The ensemble mean indicates that ~~5 and 6~~ five and six well-placed stations could explain about 75% and 80% of the SIV anomaly variance, respectively. ~~As per these numbers,~~ Adding further to six well place locations the statistical predictability does not substantially improve by adding new sites~~, taking into account that 10~~. Ten locations explain about 84% of the variance. However, as suggested by Fig. ~~??~~8, even though the SEM seems to fairly reproduce the SIV anomaly variance and, therefore, the long-term variability, it found more difficulties to reproduce the short-term variabilities.

(iv) Are the results model dependent, in particular, are they sensitive to horizontal resolution?

The results suggest that statistical predictability is affected by model resolution. Notwithstanding, the question of whether or not a finer horizontal resolution provides better statistical predictability depends on the metric used to evaluate the predictions (Section 3.2.2 and Fig. ~~??~~11). That is the case for RMSE, where the main target is to evaluate the reproducibility of the reconstructed values. It seems that an improved horizontal resolution allows a better trained statistical model so that the reconstructed values approach better to the original SIV anomaly (Fig. ~~??~~11a). On the other hand, if we look at the interannual variability, the predictors provided by numerical models with lower resolution are more able to approach the reconstructed time series to the original SIV anomaly (Fig. ~~??~~11b). ~~In this case, it is possible to argue that the low-resolution versions provide smoother time series, with less amount of short-term variability, making it easier for~~ As argued above, this study shows that model-based statistical predictability of SIV anomaly is sensible to the ~~statistical model to represent the long-term variation of SIV anomaly over time. Along the same lines,~~ model horizontal resolution. Further investigation is needed to better understand the impact of model resolution on the SIV predictability.

5 Conclusions

We envisage three main ways by which this work could support observationalists in a real-world observing system. The first is providing recommendations for optimal sampling locations. We believe that our multi-model approach provides a solid view of the sites that better represent the variability of the pan-Arctic SIV. Second, even if those regions are not taken into account for any reason (for instance, logistic, environmental harshness, strategical sampling, etc), observationalists could still take advantage of the "region of influence" concept. By doing so, they avoid deploying two or more observational platforms that would provide relatively similar information in terms of pan-Arctic SIV variability. Third, considering that observational platforms are already operational, our SEM could be trained with model outputs (with the same or other state-of-the-art AOGCMs) and so fed with observational data to project future pan-Arctic SIV variability. Within this context, we expect that this manuscript will provide recommendations for the ongoing and upcoming initiatives towards an Arctic optimal observing design.

Despite these promising results, we recognize that it might be harder to achieve skillful predictions in the ~~real world~~ real-world employing statistical tools because the actual SIV variability is likely noisier than the one described by AOGCM outputs. While model results provide an average representation of variables inside a grid cell, real-world observations would be much more heterogeneous. This issue is even more pronounced when looking at our main predictor (SIT) due to the inherent roughness and short-scale spatial heterogeneity of the real-world SIT. As consequence, this heterogeneity may be a source of uncertainties in a real observing system and more observations would be required for effectively predict the SIV anomaly. Some caution should be taken since our findings could be slightly different for other AOGCMs. A good perspective for addressing this issue is to reapply the methodology developed in this manuscript, but using all models that will be made available through the CMIP6. Also, with the sea ice depletion, some of the optimal sampling locations here suggested might ~~be in a free-ice region~~ in the future be ice free.

Finally, it is worthwhile mentioning the recent effort from the scientific community to enhance the Arctic observational system. This effort takes place through recent observational programs such as the Year Of Polar Prediction (YOPP) (?) and the ~~MOSAiC International Arctic Drift Expedition (~~Multidisciplinary drifting Observatory for the Study of Arctic Climate (MOSAiC; https://www.mosaic-expedition.org/; last access: ~~23 July 2019). Within this context, we expect that this manuscript will provide recommendations for the ongoing and upcoming initiatives towards an Arctic optimal observing design.~~ 01 March 2020).

Author contributions. LP, FM, and TF designed the science plan. DD computed the pan-Arctic sea ice area and volume. LP developed the statistical empirical model, conducted the data processing, produced the figures, analyzed the results, and wrote the manuscript based on the insights from all co-authors.

Data availability. All model outputs used in this study were made available through PRIMAVERA project (https://www.primavera-h2020.eu/; last access: 01 March 2020).

Competing interests. No competing interests are present.

Acknowledgements. The work presented in this paper has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement no. 727862: APPLICATE project (Advanced prediction in Polar regions and beyond). Leandro Ponsoni was funded by APPLICATE project until September 2019, and is now a Postdoctoral Researcher of the Fonds de la Recherche Scientifique – FNRS. François Massonnet ~~and Leandro Ponsoni are F.R.S.-FNRS PostDoc and Research Associate , respectively~~is a Research Associate of FNRS. David Docquier ~~is~~ was funded by the EU Horizon 2020 PRIMAVERA project ~~,~~ (grant agreement no. 641727) until September 2019, and is currently funded by the EU Horizon 2020 OSeaIce project, under the Marie Sklodowska-Curie grant agreement no. ~~641727.~~ 834493. Guillian Van Achter is founded by PARAMOUR project which is supported by the Excellence Of Science programme (EOS), also founded by FNRS. ~~The datasets used in this study were made available through PRIMAVERA project.~~ We thank two anonymous reviewers and the editor, Dr. Petra Heil, for their constructive suggestions and criticism.

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]~~ **[pg. 4, l. 29–30]**.

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)~~ **[pg. 23, l. 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]~~ **[pg. 26, l. 26–31]**.

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]~~ **[pg. 3, l. 33–34]**. 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]~~ **[pg. 3, l. 15]**.

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]~~ **[pg. 4, l. 21–26]**. 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]~~ **[pg. 9, l. 28–29].** Notice that the interpolated score maps (Fig. C, second column) preserve the best performance regions.

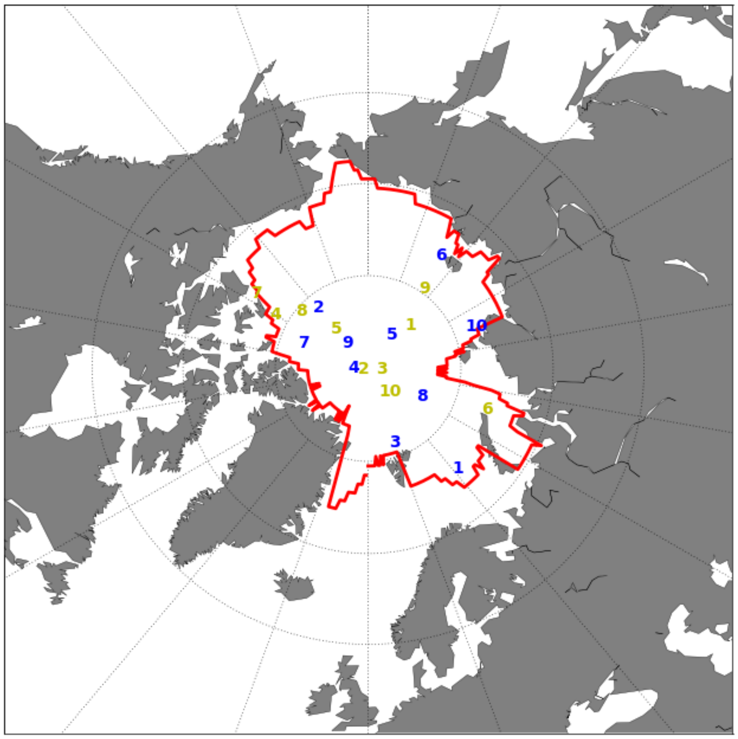

Fig. A: Map displaying two examples (out of 100) of randomly chosen locations. All random locations are placed into the area enclosed by the red line. This region represents our global region of influence as defined in Fig. 8.

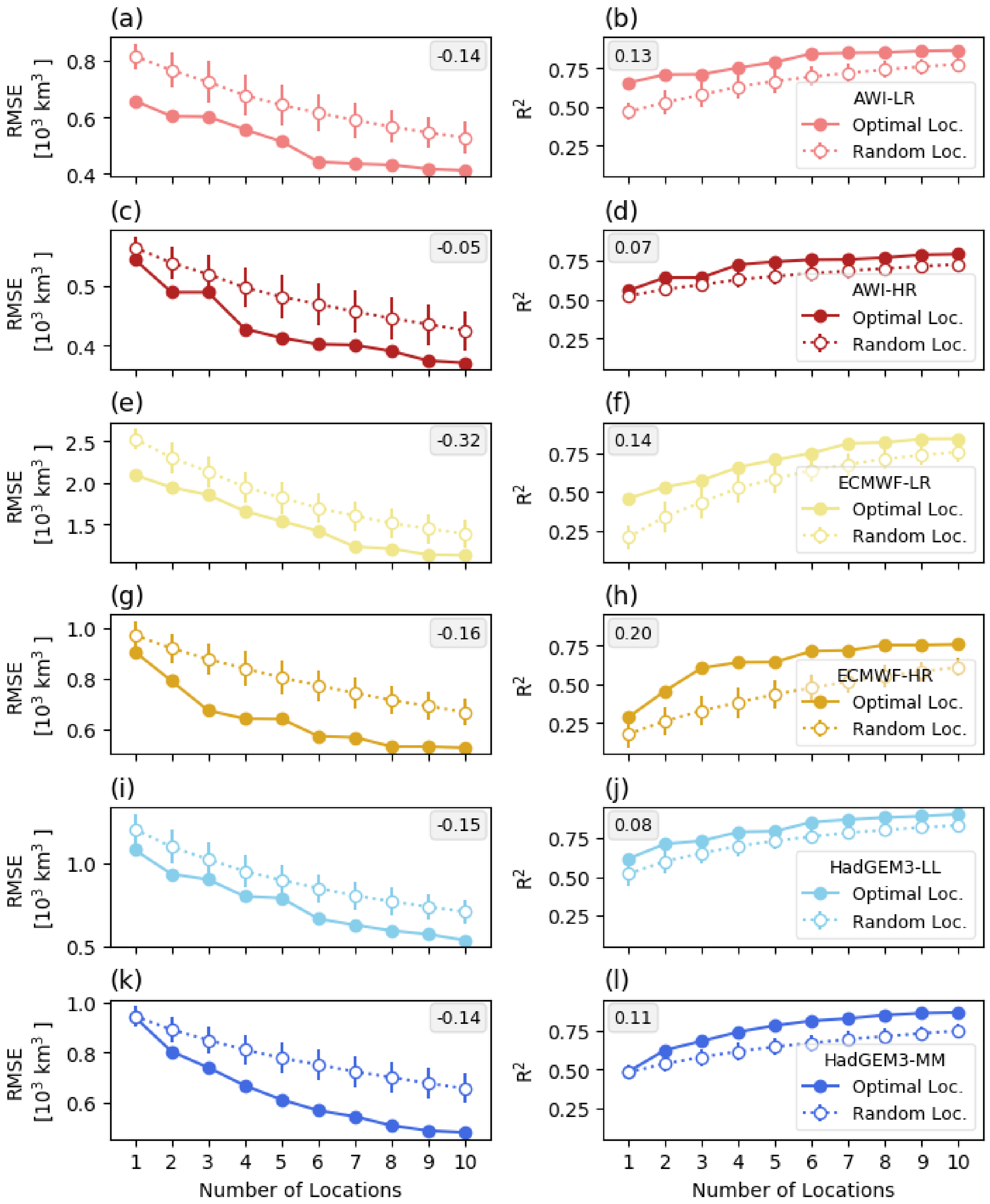

Fig. B: Root Mean Squared Error (RMSE; left column) and coefficient of determination (R^2; right column) calculated between the original and reconstructed SIV anomalies. The reconstructed SIV volume anomalies are based on the optimally selected locations following our methodology (full dots), as well 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.

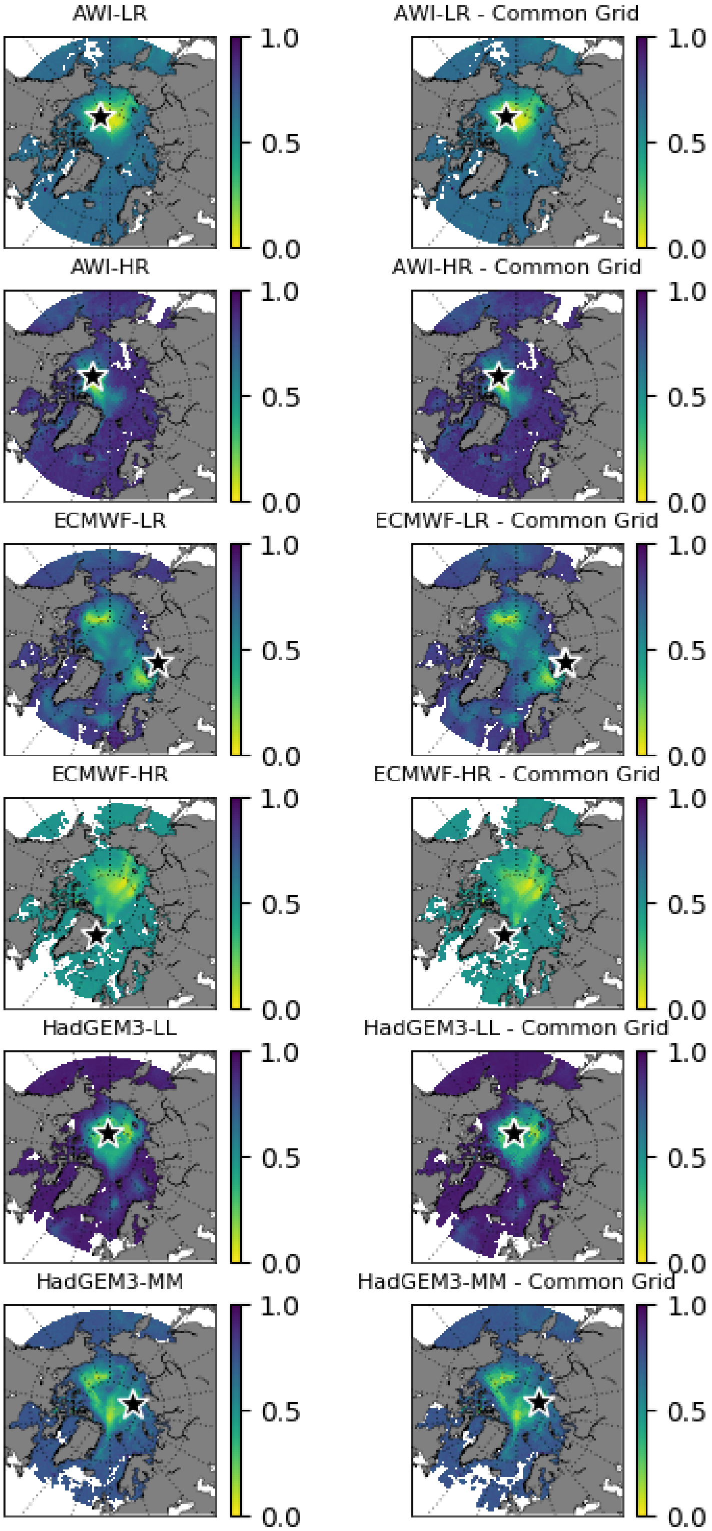

Fig. C: Normalized score maps calculated for the different model outputs with the original grid (left column) and after the interpolation to the common 1°×1° grid (right column). Notice that the interpolation has no impact on the best performance regions (shades of yellow). The interpolation is a required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

Dear Referee,

Thank you for the time that you have spent on our manuscript and for the detailed “Referee comment” report. We are happy with the 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 all 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 #2

Summary statement The motivation for this study is to contribute to an Arctic observing system by identifying key locations where sea ice thickness should be measured in order to have predictability. This study contributes to predictability and also to a stake-holder need (i.e., observationalists) of developing an efficient Arctic observing network. I feel the science is strong with really interesting (and useful) results and worthy of publication. The figures are well-prepared and understandable. My main critique is that the text needs to be smoothed out and clarified. I made detailed suggestions through about half of the document and these comments can be applied throughout the remaining parts of the paper. I have a few interpretation suggestions in the major comments. This paper is relevant for a broad science audience so the clarity of the writing is really critical for it to be broadly accessible.

Again, we thank the referee for her/his thorough review of the manuscript. We appreciated very much her/his detailed comments not only in terms of science but also regarding the writing style. Below, we answer point-by-point all major and minor comments.

Major Comments

1) It may be very useful to include stronger arguments as to why this is a model-only study. This can be strengthened in the introduction. (Page 3 lines 10-15, expand here in a way that puts it to rest). You want to be more convincing as to why this will be applicable in real life.
We have slightly changed the introduction to properly address this point. Now we have reinforced in this part of the text that observations of sea ice thickness, required for calculating the SIV, present limitations in the warmer seasons. Therefore, this variable is not made available year-round from the classical satellite campaigns ~~[pg. 3, l. 17–19]~~ **[pg. 3, l. 16–18]**. In the following paragraph ~~[pg. 3, l. 20–23]~~ **[pg. 3, l. 18–20]**, we reemphasize that the models used in this work, which are cutting edge in terms of model physics and resolution, fairly represent the thermodynamic and dynamic sea ice processes linking predictors and predictand. In Sec. ~~4~~**5**, we have added a discussion on how our study could be used in different ways by observationalists ~~[pg. 26, l. 26–33]~~ **[pg. 26, l. 17–25]**.

2) It needs to be made clear when the models are described (bottom page 3) that these are coupled climate models and are not pegged to observed conditions. Also, Discuss the GHG scenarios used for these particular simulations because all this information will make it easier for the reader to understand the results. For climate people, these are known but this paper should be accessible by weather and observational scientists as well as potentially policy experts (since they will help formulate the Arctic observing network).
That is indeed a good point. These two aspects are now clarified in the first paragraph of Sec. 2.1. ~~[pg. 4, l. 3–13]~~ **[pg. 3, l. 29 to pg. 4, l. 6]**.

3) Beginning of Section 2.2. This first paragraph lays out the methodology. I have read it twice and it is not easily understandable. Please revise this to be more precise and direct. I am not sure what to suggest specifically. Some thoughts a. Define anomaly earlier when you refer to fig 1. Just use it here. b. Move the sentence 'Overall , two categories of predictors are tested...(line 18, page 5) to be the second sentence. c. Revise the first sentence of your paragraph (your topic sentence) to something like: 'Potential predictor variables are identified for the empirical statistical model that predicts SIV anomalies.' There are extra words in this sentence and the key point of the paragraph is getting lost.
We agree with the referee. All paragraphs from Sec. 2.2 were rewritten to bring clarity to the text. To make it easier for the reader, an explanation for the term "anomaly" is provided in the Introduction ~~[pg. 2, l. 31–32]~~ **[pg. 2, l. 30]** and also in Sec. 2.1 ~~[pg. 4, 20–22]~~ **[pg. 5, 1–3]**.

4) I have some suggestions regarding the structure of the writing. a. Strengthen your 'topic sentences' that start each paragraph. This sentence should tell the reader what is in this paragraph without having to read the paragraph. The sentences in the paragraph provide the evidence or facts to support the topic sentence. This type of structure makes it easier for the reader to understand your paper quickly.
We thank the referee for the suggestion. We minutely addressed all the comments in this report taking into account this comment (4) and also the summary statement. We have promoted several changes throughout the text in order to make it clearer and easier to read for a non-specialized audience. Regarding this, we have asked for a few colleagues from different science fields to check whether or not the manuscript is understandable. Apparently, we have made the job. In any case, further comments on how to make this paper more accessible for a broader audience are always welcome.

5) It is not clear to me what the time scale for the predictions is in Section 2? (re: Fig 2, Table 1). It is one-month lead? Lag-0 is what I think it is but I did not see this explained clearly. In addition, further interpretation of the panels in Fig. 2 would be helpful because reading the 2.2 and 2.3, which refer back to Fig. 2, I see that I do not have a clear understanding or appreciation for what Fig 2 shows. It would be good to discuss each panel and provide interpretation of the panel.
It is indeed a lag-0 correlation. This is now clarified in the text ~~[pg. 6, l. 13; Fig. 2's caption]~~ **[pg. 6, l. 8; Fig. 2's caption]**. As mentioned in the answer to item (3), we have rewritten Sec. 2.2. In the new text, we are providing a better explanation of Fig. 2, considering all panels.

6) Could OHT be a poor predictor in these models because of model biases such as too strong stratification in the Arctic ocean so that 'heat' never makes it to the upper layers? This may be worthy of the discussion.
We agree with the referee. This might be a potential reason why OHT is a poor predictor. This is an interesting point that could be investigated further with more detailed analysis. We brought this discussion to the text ~~[pg. 23, l. 25–28]~~ **[pg. 23, l. 28–30]**.

7) Conclusions. The results are summarized very nicely in the model context. As an observationalist (BTW, I am a modeler), I would want to know how this is relevant in the real world. Some discussion on linking this to observations would be nice. I know this is not easy and I do not suggest that you do this research for this paper, but providing these insights will help you link it better to the people you want to use this work. If you can provide a framework that links this study to the observations, that would really strengthen the paper.

We envisage three main ways by which this work could support observationalists in a real-world observing system. The first is providing recommendations for optimal sampling locations. We believe that our multi-model approach provides a solid view of the sites that better represent the variability of the pan-Arctic SIV. Second, even if those regions are not taken into account for any reason (e.g., logistic, environmental harshness, etc), observationalists could still take advantage of the "region of influence" concept. By doing so, they avoid deploying two or more observational platforms that would provide relatively similar information in terms of pan-Arctic SIV variability. Third, considering that observational platforms are already operational, our SEM could be trained with model outputs (with the same or other state-of-the-art AOGCMs) and so fed with observational data to project future pan-Arctic SIV variability.

This discussion is now added to Sec. ~~4~~**5** ~~[pg. 26, l. 26–33]~~ **[pg. 26, l. 17–25]**.

Minor Comments

1) Page 1, Line 24, change ‘proven to bring’ to ‘led to’

Changed [pg. 2, l. 5].

2) Page 2, Line 1, change ‘disturbance of’ to ‘disturbance in’

Changed ~~[pg. 2, l. 8]~~ **[pg. 2, l. 11]**.

3) Page 2, line 1, split everything ‘which has also...’ into a separate sentence to make it easier to understand.

We have slightly reformulated the paragraph to accommodate this suggestion ~~[pg. 2, l. 5–8]~~ **[pg. 2, l. 8–12]**.

4) Page 2, line 4, change ‘sailing routes’ to ‘ship routes’, not all of the ship may be sailboats.

Changed ~~[pg. 2, l. 9]~~ **[pg. 2, l. 13]**.

5) Page 2, line 5, change ‘At global scale’ to ‘Globally’

Changed [~~pg. 2, l. 11]~~ **[pg. 5, l. 5]**.

6) Page 3, line 1, change ‘To the knowledge of the authors’ to ‘To the best of the authors’ knowledge’

Changed ~~[pg. 3, l. 4]~~ **[pg. 3, l. 3]**.

7) Page 3, line 15-16, change ‘What are the performance ..’ to ‘What is the performance...’

Changed [~~pg. 3, l. 23]~~ **[pg. 3, l. 22]**.

8) Page 3, line 17, change ‘a large amount...’ to ‘a substantial (e.g., 70%) of the original..’

We have incorporated this suggestion, but in a slightly different way. We keep the info that 70% is an arbitrarily chosen threshold ~~[pg. 3, l. 25]~~ **[pg. 3, l. 24]**.

9) Page 4, Figure 1 top panel is not even mentioned in the text. The figure panels have a and b on the right-hand side. I did not see them at first. It is standard to have them on the left corner. I suggest you edit this on all your figure panels.

We have made a proper reference to Fig. 1a. in the text ~~[pg. 4, l. 19–22]~~ **[pg. 4, l. 31–35]**. The panel index letter is now placed in the right-hand side in Fig. 1. Also in Figs. 4, 5, 11 and 12.
For Figs. 2, 6, 7 and 8 we preferred to keep the letter indicating the panel index centralized. We think the letters are easily spotted in that way.

10) Page 4, line 3, make it clear that the long-term trend and seasonal cycle has been removed. The text '(no long-term trend; no seasonal cycle)' is somewhat vague. Was there never a trend? It is clear from the top panel that there are trends but it is helpful for the reader if the language is unambiguous.
We agree with the referee. Indeed the text *"(no long-term trend; no seasonal cycle)"* is somewhat vague and confusing. We excluded this piece of text (or similar) from the entire manuscript. In the new manuscript's version, we have first defined SIV anomaly in the Introduction ~~[pg. 2, l. 31–32]~~ **[pg. 2, l. 30]**. This definition is recalled when describing Fig. 1 in Sec. 2.1 ~~[pg. 4, 20–22]~~ **[pg. 5, 1–3]**.

11) Page 5, line 7, Clarify the geographical span of the different resolution. For a student, the changing resolution is confusing.
Indeed, this is an important point. The ocean resolution of **AWI-LR** varies from **24 to 110 km**, with **∼25 km in the Arctic**. The ocean resolution of **AWI-HR** varies from **10 to 60 km**, with a refined resolution in dynamically active regions (e.g., ∼10km in the vicinity of the Gulf Stream), and **∼25 km in the Arctic**.
An important point recalled by Referee #1 is that the resolution difference between HR and LR in the Arctic is much lower in the AWI climate model compared to the other two systems. This point is clarified in the text. 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) ~~[pg. 4, l. 33–34 to pg. 5, l.1–4]~~ **[pg. 4, l. 21–26]**.

12) Section 2.3 is written very clearly. It may be worth saying something about including SIV in the SEM. IT seems to me that SIV could dominate the results since the autocorrelation is so strong in SIV.
This is a good point. We have incorporated your suggestion to the last paragraph of Sec. 2.3 ~~[pg. 7, l. 18–20 to pg. 8, l.1–6]~~ **[pg. 8, l. 24 to pg. 9, l. 2]**.

13) Section 2.4, numerous grammar issues in this section. This section is rough and needs revision.
To bring clarity to the text, we reviewed and rewrote the entire Section 2.4.

14) Fig. 10, the lag/lead time for the reconstruction is not clear to me, related to comments about Fig. 1.
As for Fig. 1, this is indeed a lag-0 comparison. This info is clarified in Sec. 3.2.2 [pg. 19, l. 7] and in the Fig. 10's caption.

15) Page 22, line 30, remove 'respectively'. I do not think that is needed here because the numbers are the same as highlighted by the word 'both'.
Indeed, "respectively" was removed from the text [pg. 25, l. 9].