Responses to Reviewer's Comments for Manuscript egusphere-2024-4092

# Improving Seasonal Arctic Sea Ice Predictions with the Combination of Machine Learning and Earth System Model

Addressed Comments for Publication to

The cryosphere

by

Zikang He, Yiguo Wang, Julien Brajard, Xidong Wang, Zheqi Shen

**Authors' Response to Reviewer #3**

**Comment 1**

The manuscript evaluates the performance of machine-learning based error correction models in a coupled earth system model, NorCPM. The evaluation compares both online and offline correction schemes. In this study, the machine learning model is trained and validated against the reanalysis generated from the same coupled earth system model, which is used as the truth. The online scheme is trained to correct instantaneous state wheares the offline scheme is used to correct monthly biases. The manuscript shows improvements using both correction schemes but the offline scheme as a post-processing method beats the online scheme. This work is interesting. However, I recommend that the manuscript should be reconsidered for publication after revision.

**Response:** We very much appreciate that the reviewer found this study interesting. We thank the reviewer for providing insightful comments that have helped to significantly improve the manuscript. We carefully addressed each concern and revised the manuscript. Below, we provide our detailed point-by-point responses to the reviewer's comments. To enhance the legibility of this response letter, all the reviewer's comments are typeset in blue boxes. Rephrased or added sentences in the revised version are indicated in a gray box.

**Major comments:**

**Comment 2**

1. In the comparison between the online and offline schemes, the online scheme is applied at 15-th of each month similar to the reanalysis system. However, the manuscript lacks the discussion on the application of the DA increment in the reanalysis system. For example, does the reanalysis system use any incremental analysis update or nudging to provide a continuous correction? The manuscript also lacks the information on the updated ocean and ice state of the reanalysis system. Are they the same as the online correction scheme? For example, there is no mention of SSS in Sect.2.2 but is corrected in the online scheme.

**Response:** We thank the reviewer very much for these comments. For clarity, we have modified the description on the reanalysis dataset used in this study (L120-L134 and L136-L138 in the manuscript):

The reanalysis is available from 1980 to 2021 with 30 ensemble members. The initial states of the reanalysis on 15 January 1980 are taken from a NorESM ensemble run integrated from 1850 to 1980 with CMIP5 historical forcings. In this reanalysis, NorCPM assimilates monthly anomalies of SST, SIC, and subsurface hydrographic profile data in the middle of each month.

From 1980 to 2002, the climatology used for anomaly-field assimilation is defined over the period 1980–2010. SST and SIC observations are from HadISST2 (Titchner and Rayner, 2014) and subsurface hydrographic profile data from EN4.2.1 (Good et al., 2013). The assimilation process contains two steps addressed in Kimmritz et al. (2019): firstly, hydrographic DA updates the ocean state (Wang et al., 2017). Subsequently, SST and SIC DA occur and update the sea ice and ocean states within the ocean mixed layer. From 2003 to 2021, the climatology utilized for anomaly-field assimilation is defined from 1982 to 2016. SST and SIC observations are from OISST (Reynolds et al., 2007) and subsurface hydrographic profile data from EN4.2.1 (Good et al., 2013). Strong-coupled DA is performed to simultaneously update the sea ice and ocean states in a single step. After each assimilation step, a post-processing step is used to ensure the physical consistency of state variables. For example, the volume of each sea ice category is proportionally adjusted based on the updated SIC

(Kimmritz et al., 2018, 2019). The other model components, such as the atmosphere and land, are dynamically adjusted through the coupler during model integration between two assimilation steps.

The online error correction approach is built from the analysis increment of the reanalysis introduced in section 2.2 (Brajard et al., 2021; Gregory et al., 2024) and sequentially applied to update the instantaneous model state in the middle of each month during prediction simulation (purple line in Figure 1), which is similar to the reanalysis system (section 2.2).

### Comment 3

These information can be useful because, in a perfect scenario, if the online correction scheme can give the same increment as the analysis increment, the online correction scheme should be able to recover the reanalysis deemed as truth here. This, of course, cannot be the case in reality, but can be useful for discussing the sources of delta RMSE. For example, lack of spatial correlation due to training on individual grid points, different strategies for correction/increment applications, or the possibility that instantaneous random errors can be averaged out so that the ML only learns systematic biases during the training processes due to the long-term data being used.

**Response:** We agree with the reviewer on this comment. For clarity, we have added discussions on the remain/unpredicted error into the manuscript (L413-L422 in the manuscript):

Furthermore, the current ML model (MLP) is trained independently at each grid point and thus cannot capture spatial correlations. This limits its ability to correct spatially coherent errors, particularly in regions where NorCPM already performs well and only subtle adjustments are needed. As a result, the hybrid model often struggles to reproduce the reanalysis, which are treated as the "truth" in this study. While it is unrealistic to expect the model to perfectly replicate analysis increments, the discrepancy is closely related to the ML-based model's learning capacity and the nature of the underlying errors. Possible contributing factors include: (1) the lack of spatial dependencies due to pointwise training and (2) the tendency of models trained on long-term data to learn systematic biases rather than instantaneous random errors, the latter of which tend to be averaged out over time. Therefore, there is still room to improve the ML-based error correction framework. Future studies could explore spatially-aware architectures, such as CNNs and U-Net, and incorporate additional predictors to capture complex error structures and enhance correction performance (Palerme et al., 2024).

### Comment 4

2. The definition of error needs to be reformulated. The analysis increment is the differences between the analysis and forecast, which is equivalent to the differences between the analysis and forecast error, xa - xf = ea - ef. Even if we take ea = 0, the increment is the negative error of the xf. Therefore, if Eq.(2) is the estimated model error, Eq.(3) means that an error is added to the model forecast. In fact, the error should be removed. The authors may want to add a negative sign to Eq.(2). The same logic can be applied to Sect.3.1 where I believe the negative error, instead of the actual error is presented.

**Response:** We agree with the reviewer that the definitions and equations in sections 2.3, as well as Figures 3 and 4 in section 3.1, are confusing. For clarity, we have revised the relevant paragraph in section 2.3 as follows (L144-L152 in the manuscript):

> The online approach is to emulate the difference between the forecast and the analysis $\mathbf{x}_k^f - \mathbf{x}_k^a$, which corresponds to the opposite of the analysis increment in DA. The error prediction model can be expressed as:
>
> $$\varepsilon = \mathcal{M}_{\mathrm{e}}(\mathbf{x}^f), \tag{1}$$
>
> where $\mathcal{M}_{\mathrm{e}}$ represents the data-driven model taking the instantaneous model state $\mathbf{x}^f$ as input and $\varepsilon$ represents the predicted model error.
> The hybrid model, incorporating the dynamic model and the online error correction model, can be expressed as follows:
>
> $$\mathbf{x}_l^h = \mathcal{M}(\mathbf{x}_{l-1}^h) - \mathcal{M}_{\mathrm{e}}(\mathcal{M}(\mathbf{x}_{l-1}^h)), \tag{2}$$
>
> where $\mathbf{x}_l^h$ represents the error-corrected instantaneous model state at $t_l$ during the prediction.
> We aim to correct SIC, SST, and SSS errors in the ice-covered area, which are directly associated with the sea ice condition.

To be consistent, we have revised Eq. (3) as follows (L150 in the manuscript):

> $$\mathbf{x}_l^h = \mathcal{M}(\mathbf{x}_{l-1}^h) - \mathcal{M}_{\mathrm{e}}(\mathcal{M}(\mathbf{x}_{l-1}^h)), \tag{3}$$

Accordingly, we have revised Figures 3 and 4 of the manuscript (Figures R1 and R2).

### Comment 5

> 3. As one of the selling point of this manuscript is the use of fully coupled ESM, can the authors provide some analysis and discussions on the ocean state as well such that one can get a better physical intuition of the results?

**Response:** We thank the reviewer for the valuable suggestion. We have analyzed SST and SSS. Taking the July initialization as an example, which is consistent with Figure 7 of the manuscript, we found that both SST and SSS exhibit clear improvements, particularly along the ice-edge region (Figures R3 and R4). We have added discussions into the manuscript (L398-L405 in the manuscript):

> By comparing the two error correction approaches, we find that the offline approach yields smaller errors than the online approach. The online error correction approach corrects instantaneous model errors only on the 15th day of the month, and the effect of this correction gradually weakens during model integration due to the accumulation of errors in the other model components. Consequently, the impact of the correction becomes less evident when computing monthly-averaged outputs. Nevertheless, the online error correction can reduce errors in SST and SSS (Figures S7 and S8). Moreover, the online correction approach maintains better physical consistency among the predicted variables through dynamical model integration. The offline error correction approach directly corrects the model outputs without requiring model integration. As a result, it is

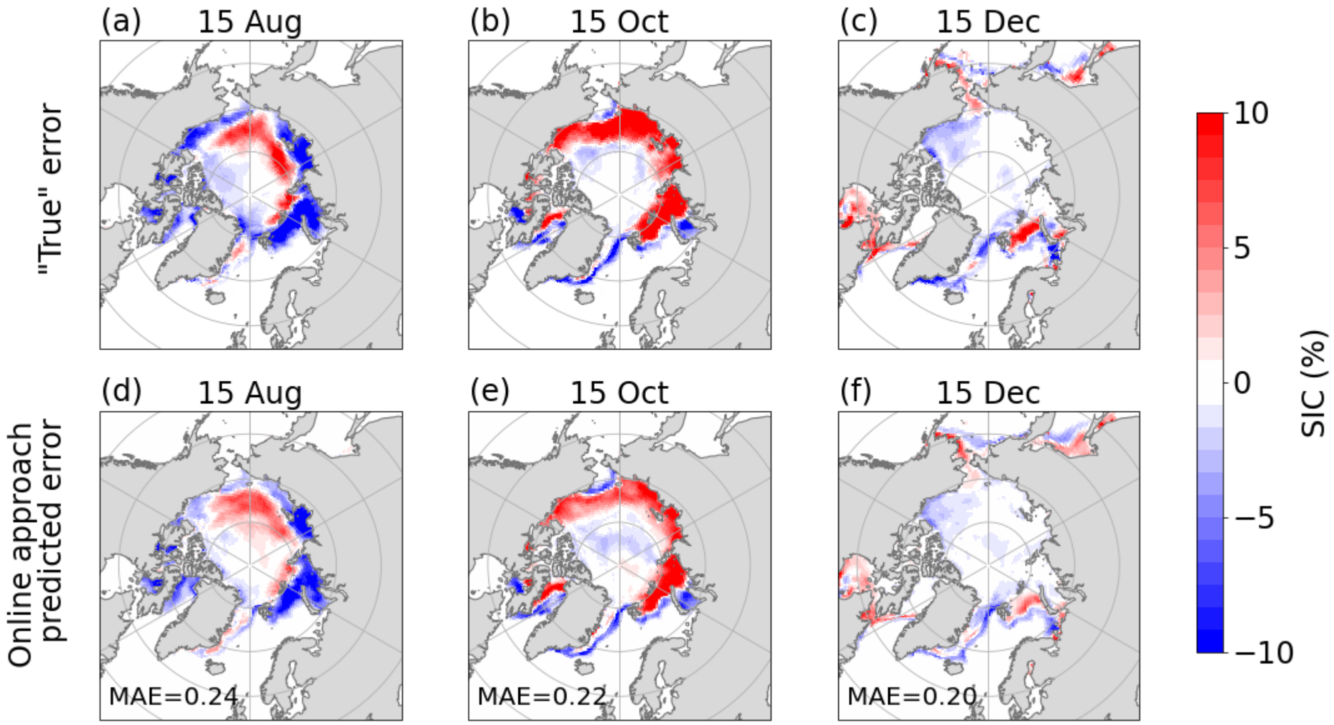

**Figure R1.** Top row: "true" errors of SIC in the middle of the month based on the analysis increments (i.e., the changes thanks to monthly DA in the reanalysis). Bottom row: the errors predicted by the online error correction model. These errors are averaged over the period 2003-2021. Values in the bottom row are the MAE between the "true" and predicted errors across space.

computationally more efficient and easier to integrate into operational sea ice prediction systems than the online approach.

**Minor comments:**

**Comment 6**

1. L28: perhaps reads better with "transitioning to DA methods to...."

**Response:** We have revised the text as follows (L31-L33 in the manuscript):

> Simultaneously, many prediction centers are transitioning to use DA methods to mitigate uncertainties in initial conditions (Wang et al., 2013; Vitart et al., 2017; Blockley and Peterson, 2018; Kimmritz et al., 2019; Wang et al., 2019; Bushuk et al., 2024).

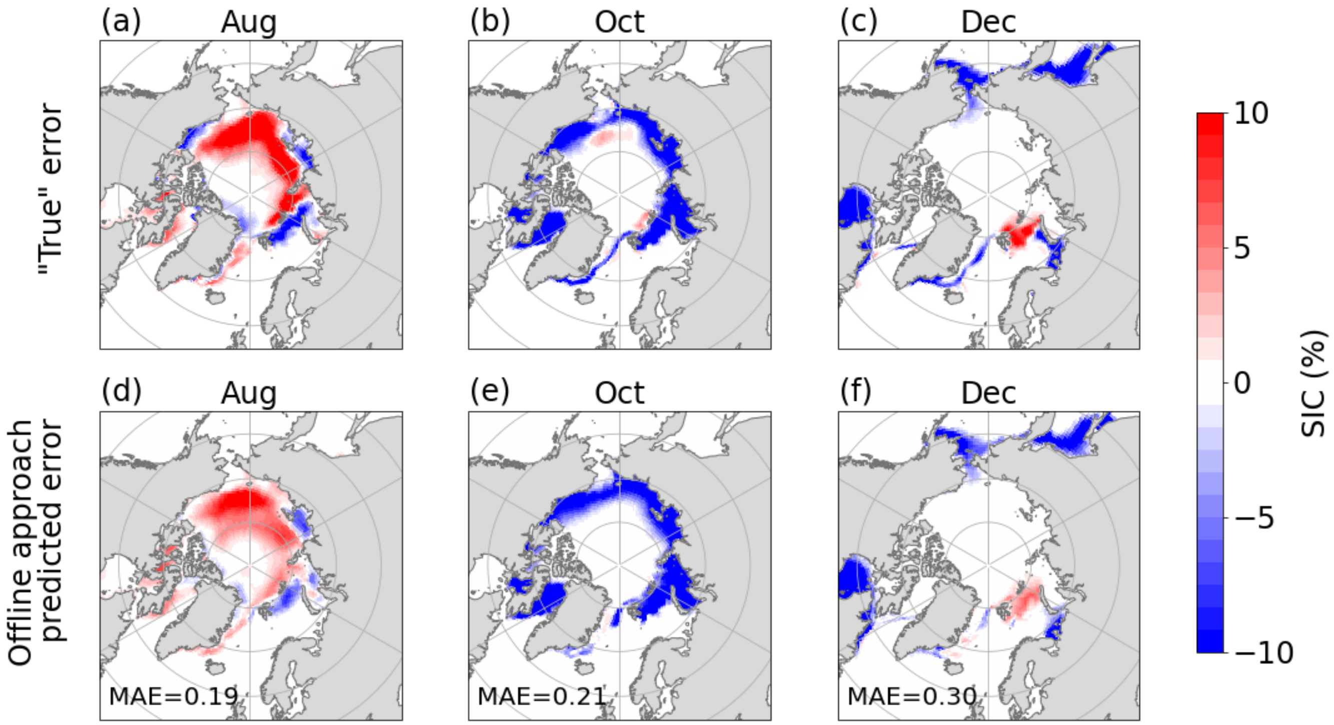

**Figure R2.** Top row: "true" errors of monthly SIC estimated by the Reference hindcast initialized in July minus the reanalysis. Bottom row: the errors predicted by the offline approach. The errors are averaged over the period 2003-2021. Values in the bottom row are the MAE between the "true" and predicted errors across space.

**Comment 7**

2. L73: Does NorCPm use DEnKF instead of stochastic EnKF? Would it be more informative to cite SAKOV, P. and OKE, P.R. (2008), A deterministic formulation of the ensemble Kalman filter: an alternative to ensemble square root filters.

**Response:** The reviewer is right. We do use DEnKF (a deterministic version of EnKF). For clarity, we have revised the manuscript as follows (L77-L78 in the manuscript):

> It combines the Norwegian Earth System Model version 1 (NorESM1, Bentsen et al., 2013) and a deterministic formulation of an advanced flow-dependent DA method named ensemble Kalman filter (EnKF, Sakov and Oke, 2008).

**Comment 8**

3. L100: I'm not sure EnKF used here actually provide a spatiotemporal estimate as normally filtering only provide spatial correlation in their error. Perhaps it is better to say "time-dependent spatial error estimate"?

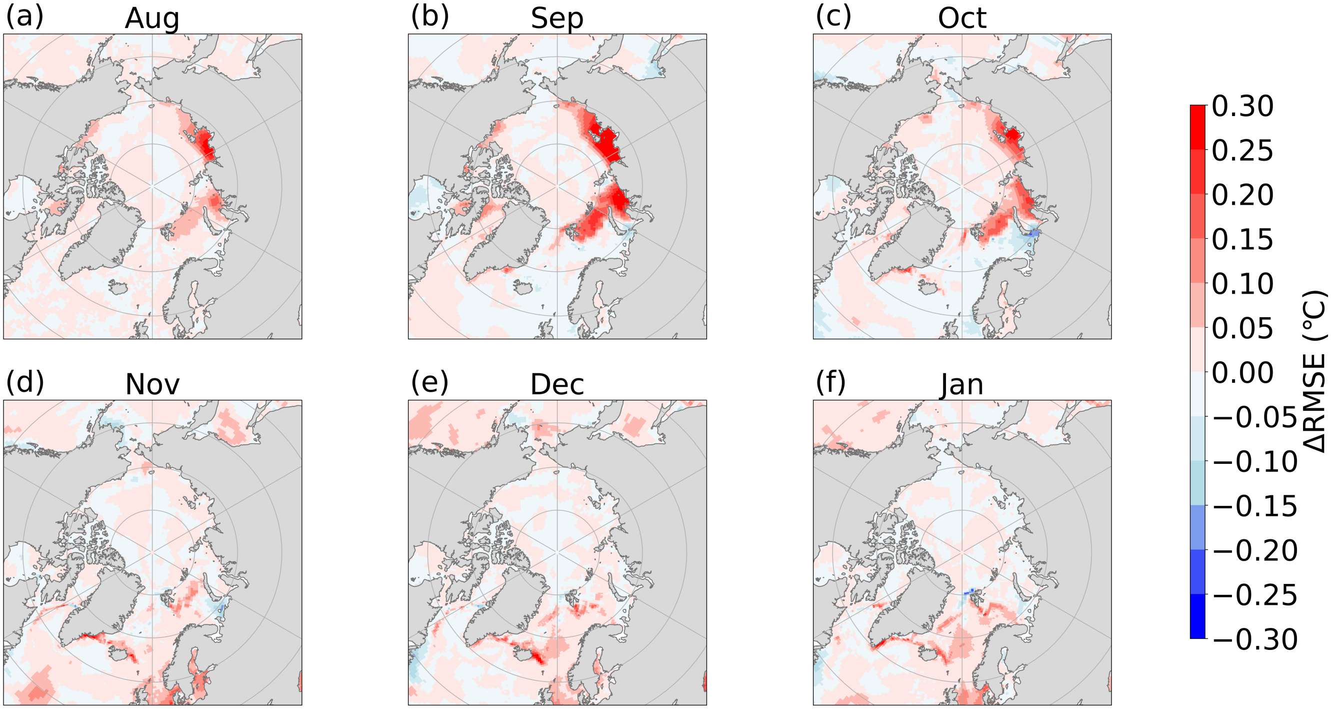

**Figure R3.** Differences between SST RMSE of the Reference and OnlineML hindcasts initialized from July. Warmer (colder) colors indicate that the OnlineML hindcast performs better (worse).

**Response:** We thank the reviewer for the comment. For clarity, we have modified the text as follows (L103-L105 in the manuscript):

> The EnKF accounts for uncertainties in initial conditions to generate ensemble predictions, which evolve in time and provide time- and space-dependent error estimates.

### Comment 9

4. L140-143: It should be explicitly said that the same post-processing of NorCPM is used in the online correction scheme.

**Response:** We have revised the manuscript by providing explicit information on the post-processing used in the study (L158-L166 in the manuscript):

> Before restarting the model after applying online error correction, it is essential to ensure that the updated variables remain within physical limits (e.g., SIC between 0% and 100%) and maintain consistency with non-updated variables. If unphysical values or inconsistencies arise, they can lead to model instability. To

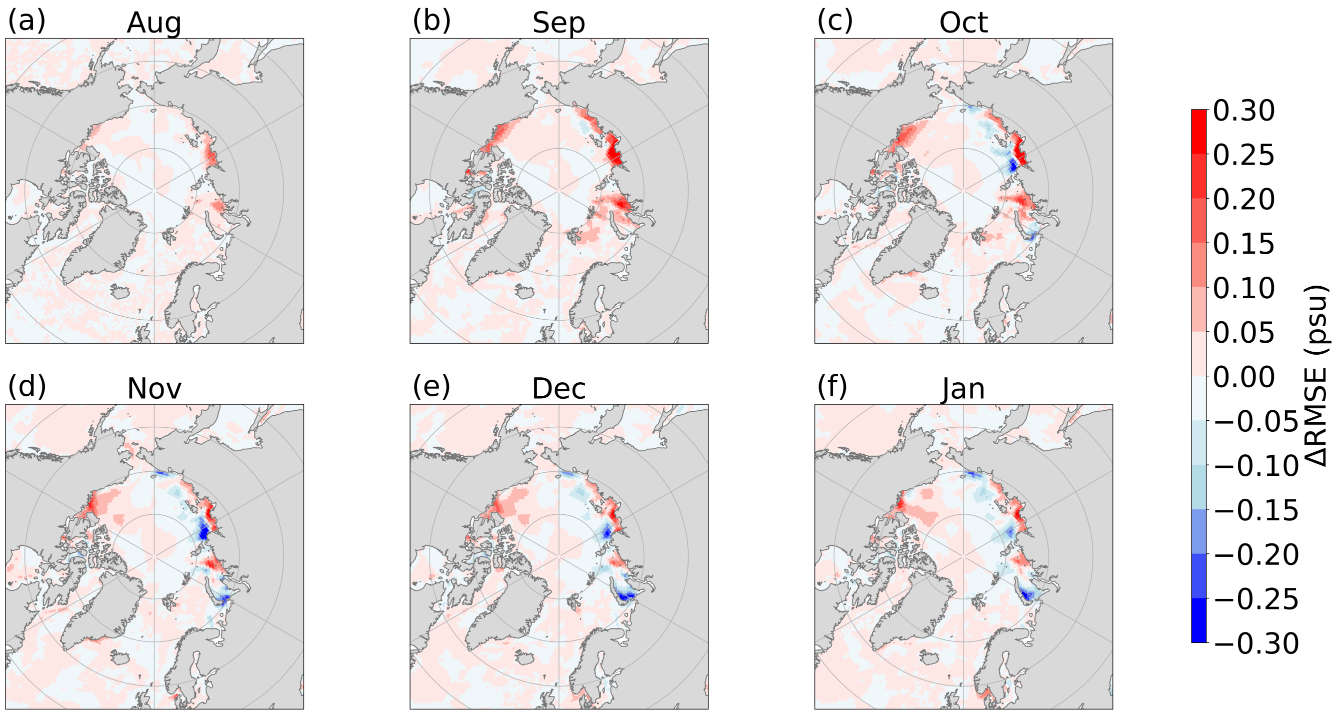

**Figure R4.** Differences between SSS RMSE of the Reference and OnlineML hindcasts initialized from July. Warmer (colder) colors indicate that the OnlineML hindcast performs better (worse).

prevent these issues, we apply a post-processing method specifically designed for NorCPM (Kimmritz et al., 2018):

- If SIC in any thickness category falls below 0% or exceeds 100%, it is set to 0% or 100%, respectively.
- If the total SIC across all thickness categories exceeds 100%, SIC values in each category are proportionally scaled to ensure the total does not surpass 100%.
- Sea ice volume in each category is adjusted proportionally to changes in SIC while preserving the ice thickness.

This approach ensures physical plausibility and model stability after error correction.

## Comment 10

5. Sect.2.4: Is post-processing applied to output from offline schemes for physical consistency when comparing online and offline schemes?

**Response:** We thank the reviewer for the question. There was a minor treatment for unphysical values in the offline scheme. For clarity, we have modified the text as follows (L169-L171 in the manuscript):

> The input features are monthly SST, SSS, total SIC, and latitude. The output feature is the error in the monthly SIC. The predicted error is subtracted from the monthly SIC. If the updated monthly SIC falls below 0% or exceeds 100%, it is set to 0% or 100%, respectively.

### Comment 11

6. Sect. 2.5: what is the objective function being used here? Is it RMSE?

**Response:** The objective function used in this study is the mean squared error (MSE). We have added the following text in the manuscript as follows (L200 in the manuscript):

> The objective function used in this study is the mean squared error (MSE).

### Comment 12

7. L193: Why is the reference configuration performed from 1991 - 2002 which is not performed for the online experiment?

**Response:** We thank the reviewer for the comment. There was a typographical error; the data actually start from 1992. In our approach, we employ a rolling training strategy, using data from the 11 years preceding each test year to train the ML models. This design ensures that no "future" information is used during training or validation. Since our test period spans 2003–2021, the first test case in 2003 requires training and validation data from 1992 to 2002. Consequently, this period could not be included in the test set. For further details, please refer to the relevant paragraph in the manuscript as follow (L210–L220 in the manuscript):

> We adopt a running training strategy, using data from the 11 years preceding the test set to train the ML models. For instance, to develop error correction models for predictions in 2011 (a test set), we train the model using data from 2000 to 2009 and validate it with data from 2010. Similarly, for predictions in 2021, we use data from 2010 to 2019 for training and data from 2020 for validation. This approach ensures that the ML models leverage the most recent data while maintaining a clear separation between training, validation, and test sets. The primary reason for using running training is the pronounced decline trend in Arctic sea ice observed over recent decades, with substantial differences between earlier ice conditions (e.g., the 1980s) and those of recent years (e.g., the 2010s). We also performed sensitivity studies on the length of the running training set (e.g., the most recent 5 years or all years since 1980) and the comparison between the running training and the fixed-period training (1992-2002), which are not shown in the paper. We found that the data from the most recent 11 years leads to the best performance for ML training, and the running training outperforms the fixed-period training.

### Comment 13

8. L207: What is an areal sum of grid points? Is it an area-weighted sum of the SIC over all grid points with SIC >= 15%?

**Response:** Sorry for the confusion. For clarity, we have revised the manuscript as follows (L247-L249 in the manuscript):

> In this study, the SIE is defined as the total area of all grid points within the region of interest where SIC $\geq$ 15%. SIE is calculated for each ensemble member, and we evaluate the ensemble mean by averaging SIE across all ensemble members.

### Comment 14

9. L226-229: What does "10 data points from the 10 ensemble members" mean? Do you mean selecting one data point from each of the 10 ensemble members leading to 10 data points in total, or do you mean selecting 10 data points from each of the 10 ensemble members leading to 100 data points in total? Based on the RMSE in Eq.(4), the RMSE is calculated over time, this means that one RMSE will be obtained for each grid point. How can a single RMSE over both time and space be obtained? Are there any results for the uncertainty of the RMSE in this manuscript?

**Response:** Sorry for the confusion. We aimed to estimate the RMSE uncertainties related to the ensemble mean. We resampled with replacement 10 data from the ensemble, ensuring the data size of 10 is equal to the ensemble size. As Eq. (5), we compute RMSE for both SIC and SIE over time and not over space (e.g., Figures 5-7 in the manuscript). For clarity, we have revised the manuscript as follows (L271-L275 in the manuscript):

> To estimate the uncertainties in an RMSE value arising from the small ensemble size, we employ the bootstrap method. Specifically, we randomly sample 10 ensemble members with replacement from the ensemble, compute the ensemble mean, and then calculate the RMSE (for either SIC or SIE) based on this resampled data. This process is repeated 10,000 times, producing a distribution of 10,000 RMSE values. The standard deviation of this distribution is then used to quantify the uncertainties associated with the RMSE value.

In Figures 5b-c of the manuscript, we used the uncertainties to perform the significant test. The ΔRMSE that does not pass the 95% significance test was marked with a black dot. In Figure 6 of the manuscript, the uncertainties of SIE RMSE were shown as error bars. In Figure 7 in the manuscript, the RMSE differences that are not statistically significant were plotted as white colors.

### Comment 15

10. L232: "...in predicting..."

**Response:** As suggested, we have revised the text as follows (L278 in the manuscript):

> We first demonstrate the performance of ML-based error correction models in predicting the model errors.

## References

Bentsen, M., Bethke, I., Debernard, J. B., Iversen, T., Kirkevåg, A., Seland, Ø., Drange, H., Roelandt, C., Seierstad, I. A., Hoose, C., et al.: The Norwegian Earth System Model, NorESM1-M–Part 1: description and basic evaluation of the physical climate, Geoscientific Model Development, 6, 687–720, 2013.

Blockley, E. W. and Peterson, K. A.: Improving Met Office seasonal predictions of Arctic sea ice using assimilation of CryoSat-2 thickness, The Cryosphere, 12, 3419–3438, 2018.

Brajard, J., Carrassi, A., Bocquet, M., and Bertino, L.: Combining data assimilation and machine learning to infer unresolved scale parametrization, Philosophical Transactions of the Royal Society A, 379, 20200 086, 2021.

Bushuk, M., Ali, S., Bailey, D. A., Bao, Q., Batté, L., Bhatt, U. S., Blanchard-Wrigglesworth, E., Blockley, E., Cawley, G., Chi, J., et al.: Predicting September Arctic Sea Ice: A Multi-Model Seasonal Skill Comparison, Bulletin of the American Meteorological Society, 2024.

Good, S. A., Martin, M. J., and Rayner, N. A.: EN4: Quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates, Journal of Geophysical Research: Oceans, 118, 6704–6716, 2013.

Gregory, W., Bushuk, M., Zhang, Y., Adcroft, A., and Zanna, L.: Machine learning for online sea ice bias correction within global ice-ocean simulations, Geophysical Research Letters, 51, e2023GL106 776, 2024.

Kimmritz, M., Counillon, F., Bitz, C., Massonnet, F., Bethke, I., and Gao, Y.: Optimising assimilation of sea ice concentration in an Earth system model with a multicategory sea ice model, Tellus A: Dynamic Meteorology and Oceanography, 70, 1–23, 2018.

Kimmritz, M., Counillon, F., Smedsrud, L. H., Bethke, I., Keenlyside, N., Ogawa, F., and Wang, Y.: Impact of ocean and sea ice initialisation on seasonal prediction skill in the Arctic, Journal of Advances in Modeling Earth Systems, 11, 4147–4166, 2019.

Palerme, C., Lavergne, T., Rusin, J., Melsom, A., Brajard, J., Kvanum, A. F., Macdonald Sørensen, A., Bertino, L., and Müller, M.: Improving short-term sea ice concentration forecasts using deep learning, The Cryosphere, 18, 2161–2176, 2024.

Reynolds, R. W., Smith, T. M., Liu, C., Chelton, D. B., Casey, K. S., and Schlax, M. G.: Daily high-resolution-blended analyses for sea surface temperature, Journal of climate, 20, 5473–5496, 2007.

Sakov, P. and Oke, P. R.: A deterministic formulation of the ensemble Kalman filter: an alternative to ensemble square root filters, Tellus A: Dynamic Meteorology and Oceanography, 60, 361–371, 2008.

Titchner, H. A. and Rayner, N. A.: The Met Office Hadley Centre sea ice and sea surface temperature data set, version 2: 1. Sea ice concentrations, Journal of Geophysical Research: Atmospheres, 119, 2864–2889, 2014.

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., Déqué, M., Ferranti, L., Fucile, E., Fuentes, M., Hendon, H., Hodgson, J., Kang, H.-S., Kumar, A., Lin, H., Liu, G., Liu, X., Malguzzi, P., Mallas, I., Manoussakis, M., Mastrangelo, D., MacLachlan, C., McLean, P., Minami, A., Mladek, R., Nakazawa, T., Najm, S., Nie, Y., Rixen, M., Robertson, A. W., Ruti, P., Sun, C., Takaya, Y., Tolstykh, M., Venuti, F., Waliser, D., Woolnough, S., Wu, T., Won, D.-J., Xiao, H., Zaripov, R., and Zhang, L.: The Subseasonal to Seasonal (S2S) Prediction Project Database, Bulletin of the American Meteorological Society, 98, 163 – 173, https://doi.org/10.1175/BAMS-D-16-0017.1, 2017.

Wang, W., Chen, M., and Kumar, A.: Seasonal prediction of Arctic sea ice extent from a coupled dynamical forecast system, Monthly Weather Review, 141, 1375–1394, 2013.

Wang, Y., Counillon, F., Bethke, I., Keenlyside, N., Bocquet, M., and Shen, M.-l.: Optimising assimilation of hydrographic profiles into isopycnal ocean models with ensemble data assimilation, Ocean Modelling, 114, 33–44, 2017.

Wang, Y., Counillon, F., Keenlyside, N., Svendsen, L., Gleixner, S., Kimmritz, M., Dai, P., and Gao, Y.: Seasonal predictions initialised by assimilating sea surface temperature observations with the EnKF, Climate Dynamics, 53, 5777–5797, 2019.