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 #2

Comment 1

The Arctic Ocean is warming at a faster rate than the rest of the planet. Consequently, both the extent and thickness of sea ice have significantly decreased over the past few decades. These changes pose significant challenges to the reliability of seasonal Arctic sea ice predictions. This manuscript aims to enhance the Norwegian Climate Prediction Model (NorCPM) for seasonal Arctic sea ice prediction by integrating machine learning techniques with the Earth system model. Online and offline modules were selected to perform error corrections. The ultimate goal is to improve NorCPM's forecast performance for the marginal ice zone.
I find this study to be both timely and relevant. It aligns well with the scope of the TC Journal. I am inclined to give a positive recommendation. However, I believe there are several issues that need to be addressed before the manuscript can possibly be considered for publication.

Response: We very much appreciate that the reviewer found this study relevant and suitable for The Cryosphere (TC). We thank the reviewer for providing insightful comments that 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, the reviewer's comments are typeset in blue boxes. Rephrased or added statements in the revised version of the manuscript are indicated in gray boxes.

Comment 2

(1) L158: "MLP excels in function approximation, making it particularly..." Please explain why the MLP (Multilayer Perceptron) model was chosen over the convolutional neural network (CNN). In my opinion, a CNN is better suited for learning spatial neighboring relationships compared to an MLP model. However, using an MLP model to operate on each grid point could lead to abrupt spatial changes in the predicted values.

Response: We sincerely appreciate the reviewer's comment. We explained the choice of MLP in the previous version of the manuscript. For clarity, we have revised the manuscript as follows (L183-L188 in the paper):

> The ML architecture used in this study is a multilayer perceptron (MLP), a fully connected neural network well-suited for capturing complex nonlinear relationships in data. MLP offers several advantages, including flexibility in handling diverse input features, efficient training via backpropagation, and strong generalization when properly regularized. Additionally, MLP is computationally more efficient than complex deep learning architectures such as convolutional neural networks (CNNs) and U-Net. It has been successfully applied to error correction in geophysical modeling (e.g., Yang et al., 2023), as it is computationally efficient and requires less training data (Jia et al., 2019; Watson, 2019).

Table R1. Number of parameters of the online and offline ML-based error correction models for each ML model.

	Online ML-based SIC model	Online ML-based SST/SSS model	Offline ML-based SIC model
BatchNorm	52	52	20
Dense layer 1	840	840	360
Dense layer 2	1830	1830	1830
Gate layer	31	31	31
Dense layer 3	840	840	360
Dense layer 4	1830	1830	1830
Output	155	31	31

Comment 3

(2) L171: The attention mechanism is a widely used deep learning technique. I suggest that the authors provide a detailed explanation of the attention mechanism employed in this study. Was a temporal attention mechanism or a self-attention mechanism used in this context? Generally, the number of parameters in an attention module is significantly larger than that in an MLP model. Given the volume of data in this study, there is a risk of overfitting. I would like to see the parameter counts for both the MLP model and the attention mechanism presented separately.

Response: In this study, the attention mechanism employed is a gate-based attention layer, which serves to adaptively reweight the input features. Detailed parameters of each machine learning model, including the gate layer, are provided in Table R1 (i.e., Table 2 in the revised manuscript).

To address the potential risk of overfitting, we have incorporated several regularization strategies, such as early stopping, L2 regularization, and dropout. The corresponding descriptions have been added to the manuscript as follows (L200-L208 in the paper):

> The objective function used in this study is the mean squared error (MSE). Additionally, details regarding the number of parameters for each ML model are provided in Table 2. To reduce the risk of overfitting and improve model generalization, the following strategies are implemented:
>
> - **Batch Normalization**: The inputs of each layer are normalized to reduce internal covariate shift, thus promoting training stability and generalization.
> - **L2 Regularization**: A penalty is applied to the output layer weights, effectively discouraging over-complex models and reducing the likelihood of overfitting.
> - **Early Stopping**: The validation loss is monitored during training and the training is halted once the validation loss curve does not decline, avoiding overfitting due to the training data.

Comment 4

(3) L184: Training a separate model for each month of the test period (2003 to 2022) is not a convincing design. A model trained for each month during the training period should be generalizable to the test period. In my opinion, there is no need to train a distinct model for each month of each year during the test period.

Response: We respectively disagree with the reviewer. As shown in Figure R1, the running training results in lower MAE than the fixed-period training (1992 - 2002). Therefore, we adopted running training in this study. For clarity, we have included this comparison into the manuscript as follows (L210-L220 in the paper):

> 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.

Other Comments

Comment 5

The manuscript suffers some unclarities concerning the structure.

Response: We have carefully addressed each comment of the three reviewers and revised the manuscript. We believe the manuscript has been improved.

Comment 6

(4) L140: This section introduces the limitations of online error correction. However, its placement in the methods description section feels abrupt and lacks strong contextual relevance. I suggest moving this part to the discussion section, where it can be framed as a prospect for future work.

Response: Sorry for the confusion. This paragraph describes the post-processing after applying the online error correction approach. For clarity, we have revised the manuscript as follows (L158-L166 in the paper):

> 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 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%.

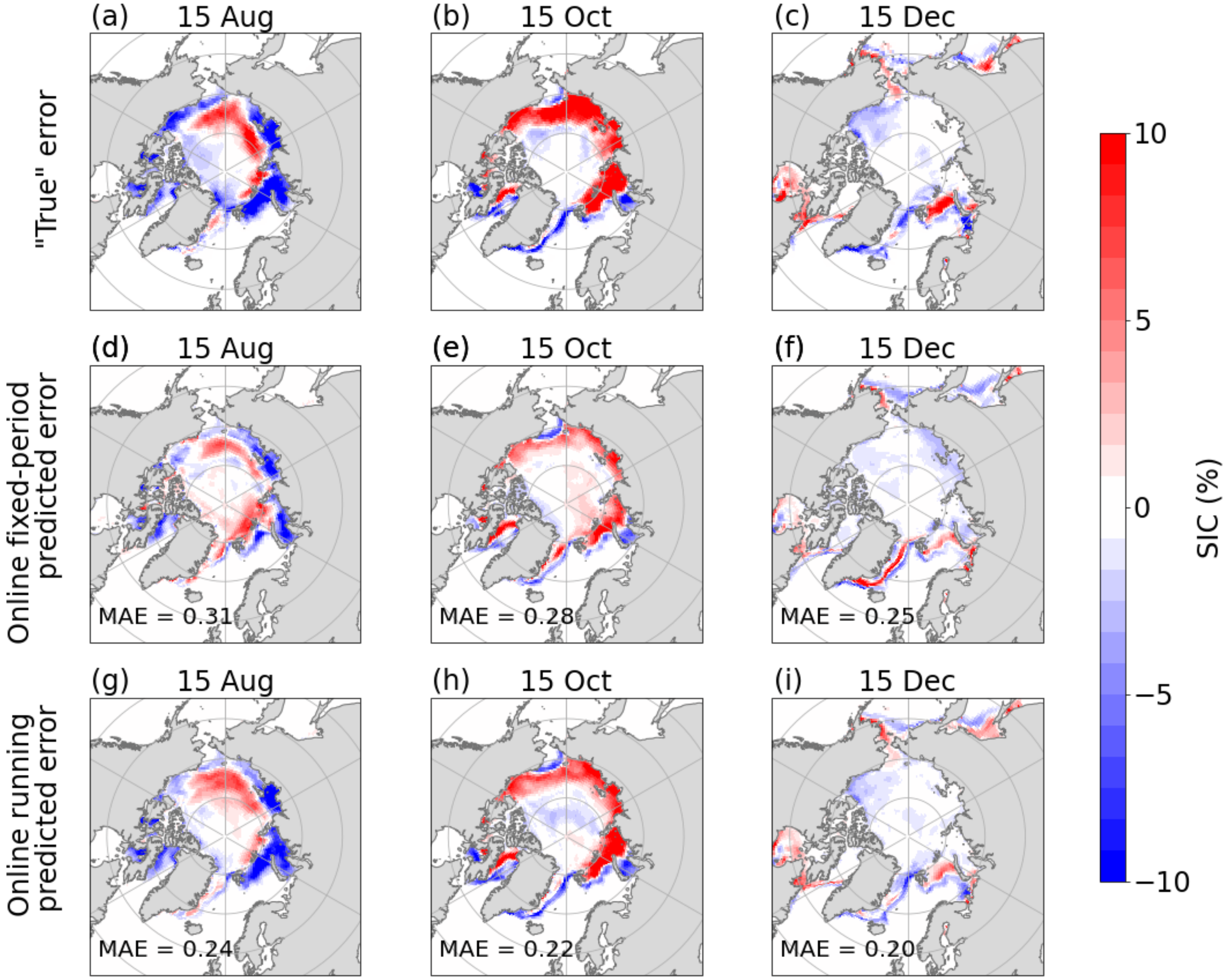

Figure R1. Top row: "true" errors of SIC in the middle of the month based on the analysis increments (i.e., the changes thanks to monthly DA in the reanalysis). Middle row: the errors predicted by the fix-period training online error correction model (1992 - 2002). Bottom row: the errors predicted by 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.

- Sea ice volume in each category is adjusted proportionally to changes in SIC while preserving the ice thickness.

This approach ensures physical constraint and model stability after the error correction.

Comment 7

(5) L165: "The MLP architecture consists of five layers" Please consider presenting this with a diagram for better clarity and algorithm flow.

Response: We thank the reviewer for the comment. Our model is relatively simple, and the study involves three different input/output configurations. To avoid potential confusion for readers, we decided not to include a schematic diagram. Instead, we have revised the textual descriptions in the manuscript to clarify the model structures (both ML configuration and the number of parameters) as follows (L189-L208 in the manuscript):

The entire MLP architecture consists of seven layers:

- **Input layer**: A batch normalization layer (Ioffe, 2017), which helps stabilize and accelerate the training process by normalizing the input features.
- **Second layer**: A dense layer with 60 neurons, using the rectified linear unit (ReLU) activation function.
- **Third layer**: A dense layer with 30 neurons, also employing the ReLU activation function. This layer shares the same structure as the second layer.
- **Fourth layer**: An attention mechanism implemented via a gate layer, which enables the model to focus on important features, thereby enhancing learning efficiency and predictive performance.
- **Fifth layer**: A dense layer with 60 neurons and ReLU activation, mirroring the configuration of the second layer.
- **Sixth layer**: A dense layer with 30 neurons and ReLU activation, identical to the third layer.
- **Output layer**: A dense layer activated by the linear function.

The objective function used in this study is the mean squared error (MSE). Additionally, details regarding the number of parameters for each ML model are provided in Table 2. To reduce the risk of overfitting and improve model generalization, the following strategies are implemented:

- **Batch Normalization**: The inputs of each layer are normalized to reduce internal covariate shift, thus promoting training stability and generalization.
- **L2 Regularization**: A penalty is applied to the output layer weights, effectively discouraging over-complex models and reducing the likelihood of overfitting.
- **Early Stopping**: The validation loss is monitored during training and the training is halted once the validation loss curve does not decline, avoiding overfitting due to the training data.

Comment 8

(6) L172: The specific use of the "linear activation function" should be clarified. Did the authors apply an activation mechanism, or was it unnecessary? These details are critical for understanding the implementation.

Response: Sorry for the confusion. The output layer is activated by the linear function, which applies no nonlinear transformation to the output. This choice is deliberate, as the task involves regression and requires unbounded continuous output values. Therefore, no activation mechanism (e.g., ReLU or sigmoid) is applied in this layer. We have modified the description as follows (L189-L199 in the manuscript):

> The entire MLP architecture consists of seven layers:
>
> – **Input layer**: A batch normalization layer (Ioffe, 2017), which helps stabilize and accelerate the training process by normalizing the input features.
>
> – **Second layer**: A dense layer with 60 neurons, using the rectified linear unit (ReLU) activation function.
>
> – **Third layer**: A dense layer with 30 neurons, also employing the ReLU activation function. This layer shares the same structure as the second layer.
>
> – **Fourth layer**: An attention mechanism implemented via a gate layer, which enables the model to focus on important features, thereby enhancing learning efficiency and predictive performance.
>
> – **Fifth layer**: A dense layer with 60 neurons and ReLU activation, mirroring the configuration of the second layer.
>
> – **Sixth layer**: A dense layer with 30 neurons and ReLU activation, identical to the third layer.
>
> – **Output layer**: A dense layer activated by the linear function.

Comment 9

(7) L175: Please provide a detailed split of the datasets. It is essential to ensure that there is no overlap in time or data between the training set, validation set, and test set to prevent data leakage. Such overlap could lead to an unreliable evaluation of the model's performance on the test set.

Response: We agree with the reviewer on the importance of avoiding overlap between the training, validation, and test sets to prevent data leakage. In our approaches, we implement a running training strategy, and there is no overlap between these sets. To enhance clarity, we have revised the relevant paragraph as follows (L210-L220 in the paper):

> 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

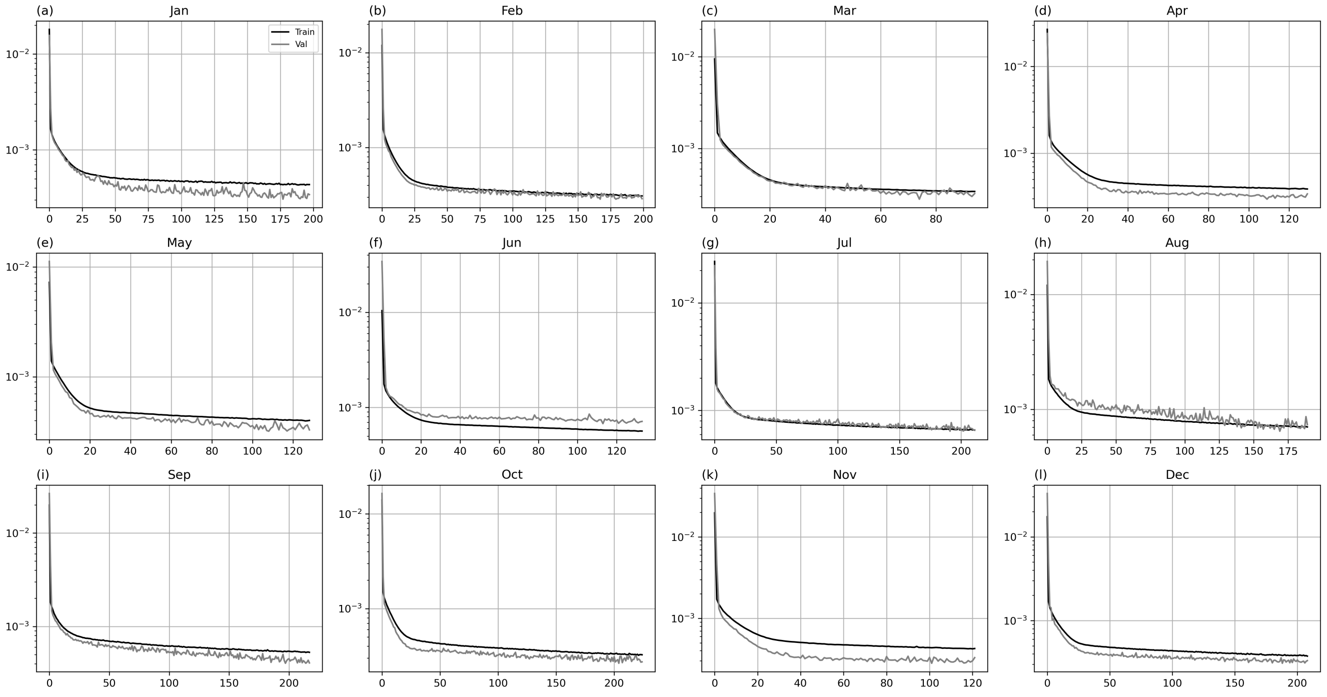

Figure R2. Training and validation loss curves of the online error correction models for the test year 2003.

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 10

(8) L230: It is necessary to evaluate the performance of the MLP model, such as giving specific training set accuracy, validation set accuracy, and test set accuracy, so as to demonstrate the generalization ability of the model and make the subsequent evaluation of specific correction effects more credible.

Response: We thank the reviewer for the comment. We have carefully examined the training curves. Considering the large number of ML models trained, we present the results for the online error correction models in the year 2003 as a case study. As shown in Figure R2, the training curves for all twelve months demonstrate satisfactory convergence. Owing to the use of early stopping, the number of training epochs varies among the different months.

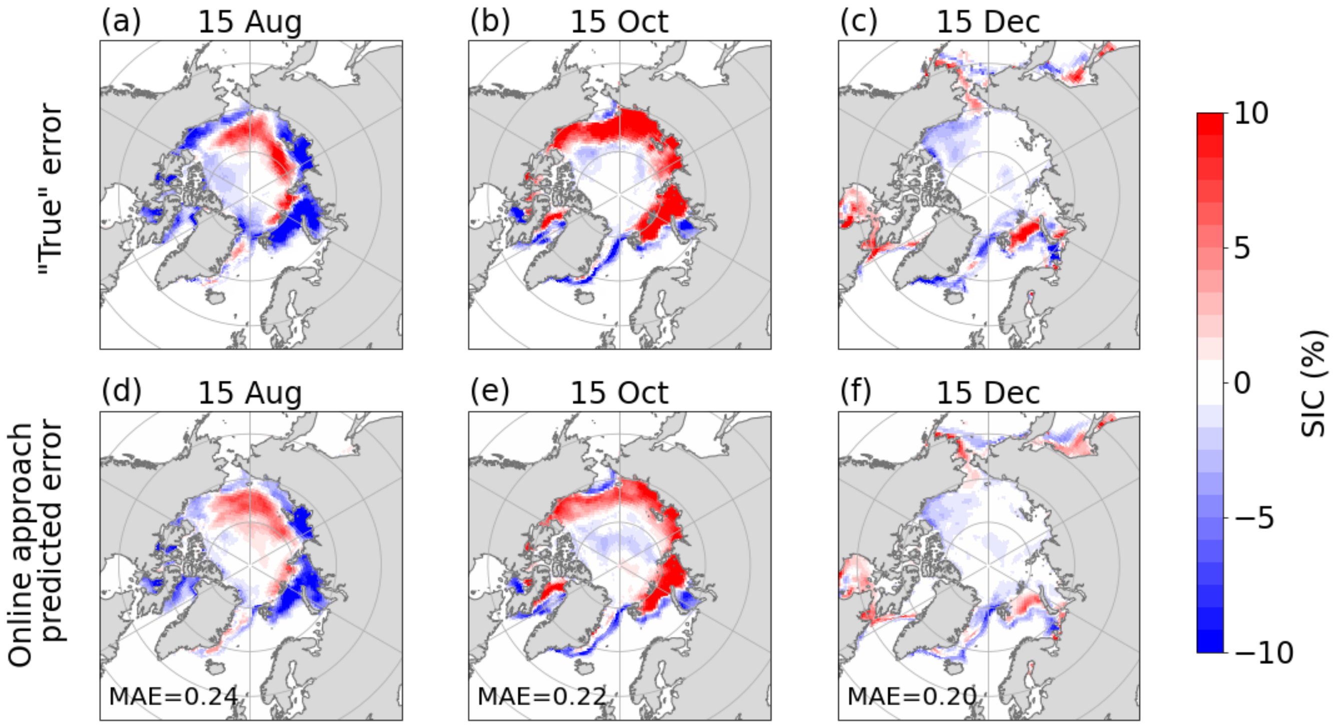

Figure R3. Top row: "true" errors of SIC in the middle of the month based on the analysis increments (i.e., the changes thanks to monthly DA in the reanalysis). Bottom row: the errors predicted by the online error correction model. These errors are averaged over the period 2003-2021. Values in the bottom row are the MAE between the "true" and predicted errors across space.

Comment 11

(9) F3, 4, 7: Please add specific accuracy or error in each subgraph. Present true errors with comma “true errors”

Response: As suggested, we have revised these figures by adding a specific accuracy (spatial mean absolute error or ΔRMSE) to each subgraph and introducing a comma for "true" or "truth" in the whole manuscript. Please refer to Figures R3 and R4 and the manuscript.

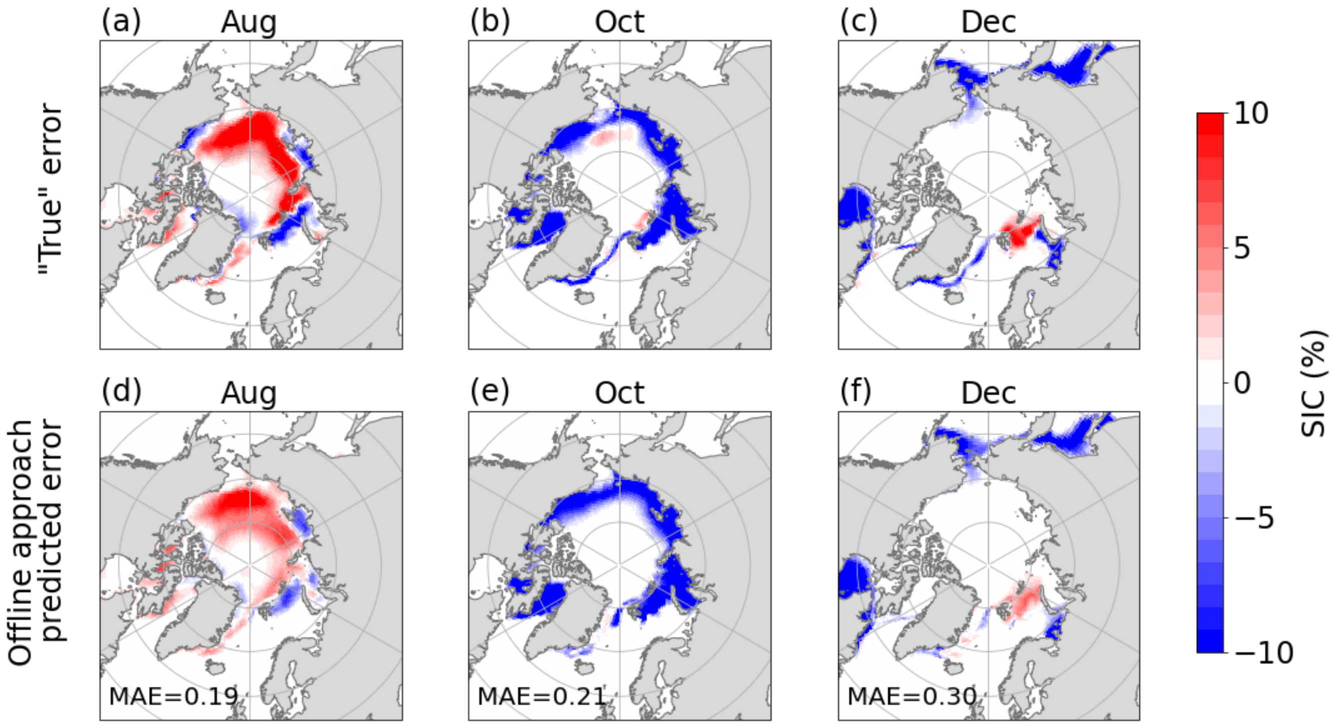

Figure R4. Top row: "true" errors of monthly SIC estimated by the Reference hindcast initialized in July minus the reanalysis. Bottom row: the errors predicted by the offline approach. The errors are averaged over the period 2003-2021. Values in the bottom row are the MAE between the "true" and predicted errors across space.

Comment 12

(10) L261: "Compared with the OnlineML hindcasts, the OfflineML hindcasts have a larger error reduction, particularly in September"; L269: "demonstrates larger error reductions in IIEE than the online approach..."; L273: "The offline approach outperforms the online approach in reducing both RMSE for SIE and IIEE for ice edge, especially in months with higher prediction errors"
The manuscript contains numerous vague expressions. Please provide more specific and concrete details. For example, instead of stating that "the error has been reduced," specify by how much (e.g., "the error has been reduced by xx%"). Including precise values of accuracy is essential for a comprehensive evaluation.

Response: We appreciate the reviewer's comment. As suggested, we have revised the manuscript by providing concrete details as follows (L309-L310, L317-L320, L321-L325, L279-L285 and L369-L377 in the manuscript):

> Compared to the OnlineML hindcast, the OfflineML hindcast achieves a greater error reduction, particularly in September, where they reduce the SIE prediction error by up to 75% relative to the Reference hindcast.

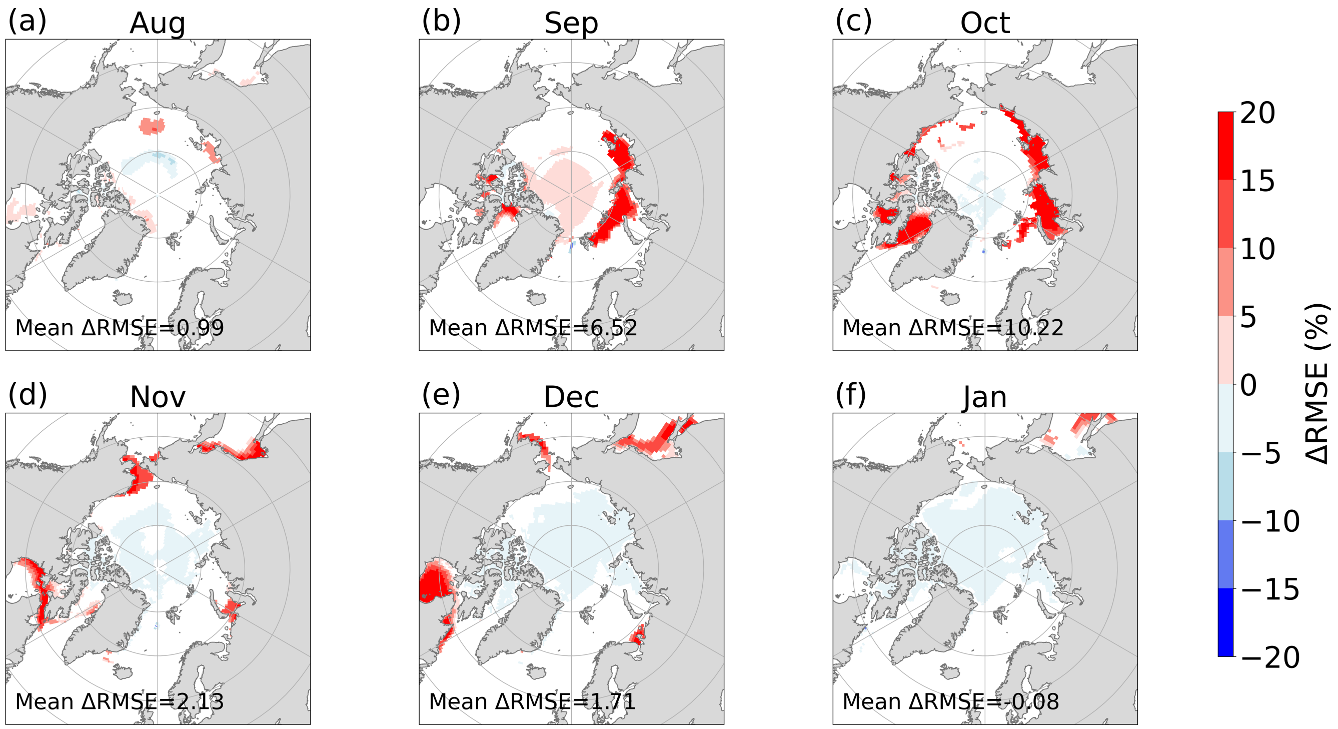

Figure R5. Differences in SIC RMSE between the Reference and OfflineML hindcasts initialized from July. Warmer (colder) colors indicate that the OfflineML hindcast outperforms (underperforms) the Reference hindcast. White areas indicate differences that are not statistically significant. The value in each subplot represents the average ΔRMSE for statistically significant grid points.

In contrast, the offline approach consistently improves performance across nearly all periods and achieves larger error reductions in IIEE than the online approach, particularly from June to January, with the maximum reduction exceeding 0.5×10^6 km^2 compared to the Reference hindcast. By directly correcting monthly mean outputs, the offline approach avoids information loss during the model integration, leading to larger error reduction.

In summary, the Reference hindcast exhibits relatively larger prediction errors from August to October, primarily due to increased model uncertainties associated with atmospheric forcing and sea ice processes. The offline approach outperforms the online approach in reducing both the RMSE of SIE and the IIEE along the ice edge, particularly during high-error months. For example, in September, the RMSE of SIE is reduced by 75%, and the IIEE is reduced by over 0.5×10^6 km^2 compared to the Reference hindcast.

The "true" errors obtained from analysis increments and the errors predicted by the online error correction model are averaged over 2003-2021 and displayed in Figure 3. The spatial patterns of the "true" error vary significantly across different dates. For instance, on August 15, errors are predominantly negative across most regions as NorCPM underestimates SIC in this month, with some localized positive errors occurring internally.

The average MAE across ice-covered grid points is 0.24%. On October 15, the errors are mostly positive as NorCPM overestimates SIC, resulting in an average MAE of 0.22%. On December 15, the MAE is 0.20%, primarily appearing in marginal ice areas, with overall lower magnitudes compared to August and October. Notably, the average error remains below 1% in all cases.

The impact of the error correction on SIC is more pronounced near the ice edge (Figure 7). In August, only a few grid points in the Siberian and Atlantic regions exhibit improvements (Figure 7a), with an average improvement of 0.99%. In September, notable enhancements appear across the central Arctic, Atlantic, Siberian, and Canadian regions (Figure 7b), with an average improvement of 6.52%. In October, significant improvements are observed in the Atlantic and Canadian regions, reaching an average of 10.22%. Additionally, some blue areas emerge in the central Arctic, indicating substantial differences between the Reference hindcast and the OfflineML hindcast, though the magnitude of these differences remains minimal. In November and December, the positive impact of the error correction is primarily concentrated in Hudson Bay and the Sea of Okhotsk. However, noise increases in the central Arctic, and the average improvement declines to 2.13% in November and 1.71% in December. The widespread presence of blue in the central Arctic also results in an average improvement of −0.08% in January.

References

Ioffe, S.: Batch renormalization: Towards reducing minibatch dependence in batch-normalized models, Advances in neural information processing systems, 30, 2017.

Jia, X., Willard, J., Karpatne, A., Read, J., Zwart, J., Steinbach, M., and Kumar, V.: Physics guided RNNs for modeling dynamical systems: A case study in simulating lake temperature profiles, in: Proceedings of the 2019 SIAM international conference on data mining, pp. 558–566, SIAM, 2019.

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.

Watson, P. A.: Applying machine learning to improve simulations of a chaotic dynamical system using empirical error correction, Journal of Advances in Modeling Earth Systems, 11, 1402–1417, 2019.

Yang, Z., Liu, J., Yang, C.-Y., and Hu, Y.: Correcting nonstationary sea surface temperature bias in NCEP CFSv2 using Ensemble-based Neural Networks, Journal of Atmospheric and Oceanic Technology, 40, 885–896, 2023.