~~Improving~~ Correcting Errors in Seasonal Arctic Sea Ice ~~Predictions with the Combination~~ Prediction of ~~Machine Learning and~~ Earth System Model with Machine Learning

Zikang He[1,2,3], Yiguo Wang[3], Julien Brajard[3], Xidong Wang[1,2], and Zheqi Shen[1,2]

[1]Key Laboratory of Marine Hazards Forecasting, Ministry of Natural Resources, Hohai University, Nanjing, 210098, China

[2]College of Oceanography, Hohai University, Nanjing, 210098, China

[3]Nansen Environmental and Remote Sensing Center and Bjerknes Centre for Climate Research, Jahnebakken 3, Bergen, N-5007, Norway

Correspondence: Yiguo Wang (Yiguo.wang@nersc.no) and Xidong Wang (xidong_wang@hhu.edu.cn)

Abstract. While ~~dynamical~~ Earth system models are essential for seasonal Arctic sea ice prediction, they often exhibit significant errors that are challenging to correct. In this study, we integrate a multilayer perceptron (MLP) machine learning (ML) model into the Norwegian Climate Prediction Model (NorCPM) to improve seasonal sea ice predictions. We compare the online and offline error correction approaches. In the online approach, ML corrects errors in the model's instantaneous state during the model simulation, while in the offline approach, ML post-processes and calibrates predictions after the model simulation. Our results show that the ML models effectively learn and correct dynamical model errors in both ~~methods~~approaches, leading to improved predictions of Arctic sea ice during ~~test periods~~ the test period (i.e., 2003-2021). Both ~~methods~~ approaches yield the most significant improvements in the marginal ice zone, where error reductions in sea ice concentration exceed ~~20%~~20%. These improvements vary seasonally, with the most substantial enhancements occurring in the Atlantic, Siberian, and Pacific regions from September to January. The offline error correction approach consistently outperforms the online error correction approach. This is primarily because the online approach targets only instantaneous model errors on the 15th of each month, while errors can grow during the subsequent one-month model integration due to interactions among the model components, damping the error correction in monthly averages. Notably, in September, the online approach reduces the error of the ~~pan-Arctic~~ Pan-Arctic sea ice extent by ~~50%~~50%, while the offline approach achieves a ~~75%~~ 75% error reduction.

1 Introduction

According to satellite observations, the Arctic sea ice extent (SIE) rapidly declines throughout all calendar months during the recent decades ~~(e.g., Serreze et al., 2007)~~(e.g., Serreze et al., 2007; Onarheim et al., 2018; Wang et al., 2022; Heuzé and Jahn, 2024). The most significant reductions occurred in the summer and autumn (e.g., September, Stroeve et al., 2014). The wider open ocean leads to growing socioeconomic activities in the Arctic (e.g., fisheries, shipping, and resource extraction). These increased human activities highly demand accurate seasonal predictions of Arctic sea ice conditions (Jung et al., 2016; Wagner et al., 2020). The Sea Ice Outlook, managed by the Sea Ice Prediction Network, produces monthly reports during the Arctic sea ice retreat season. These monthly reports synthesize input from the international research community devoted to enhancing sea ice predictions. Recently, Bushuk et al. (2024) evaluated ~~and compared~~ 17 statistical models, 17 dynamical models, and 1 heuristic approach in predicting September Arctic sea ice. They found that dynamical and statistical models are overall comparable in predicting the Pan-Arctic SIE, and dynamical models generally outperform statistical models in predicting the regional SIE and sea ice concentration (SIC, i.e., local quantities). Bushuk et al. (2024) also suggested that the dynamical models must further improve their initialization and model resolution to reduce prediction errors.

Data assimilation (DA) integrates observations with dynamical models to optimally estimate the state of the climate system ~~(Carrassi et al., 2018; Penny and Hamill, 2017)~~(Penny and Hamill, 2017; Carrassi et al., 2018). It has widespread application in producing reanalysis ~~(Saha et al., 2006; Dee et al., 2011; Zuo et al., 2019; Laloyaux et al., 2018; Hersbach et al., 2020)~~ (Saha et al., 2006; Dee et al., 2011; Laloyaux et al., 2018; Zuo et al., 2019; Hersbach et al., 2020), offering comprehensive, continuous, and dynamically consistent reconstructions of past climate states. Simultaneously, many prediction centers are transitioning to ~~DA adoption~~ use DA methods to mitigate uncertainties in initial conditions ~~(Kimmritz et al., 2019; Blockley and Peterson, 2018;~~ (Wang et al., 2013; Vitart et al., 2017; Blockley and Peterson, 2018; Kimmritz et al., 2019; Wang et al., 2019; Bushuk et al., 2024). The improved ~~density~~ quantity and quality of observations across different climate system components and advanced DA methods enable more precise initial conditions for seasonal predictions of Arctic sea ice. Nevertheless, even with perfect initial conditions, prediction errors escalate over time due to the inherent deficiencies of dynamical models in emulating the true climate system (gray ~~line and pink line~~ and green lines in Figure 1). This underscores the necessity for dealing with prediction errors.

Machine learning (ML) has recently emerged as a data-driven technique to mitigate dynamical prediction errors. Two prevalent approaches include constructing an ML-dynamical hybrid model (e.g., Brajard et al., 2021; Watt-Meyer et al., 2021) and post-processing/calibrating model ~~output (e.g. Palerme et al., 2024; Yang et al., 2023)~~outputs (e.g., Yang et al., 2023; Palerme et al., 2024). The former is considered as online error correction, while the latter refers to offline error correction.

In the context of the online error correction, ML is applied to correct errors in the instantaneous model state (i.e., initial conditions for the following model integration) and sequentially applied to update the instantaneous model state during simulation (e.g., Brajard et al., 2021), referring to an ML-dynamical hybrid model (purple line in Figure 1). Such online error correction approaches have been investigated in both an idealized framework (e.g., Watson, 2019; Brajard et al., 2021) and real applications (e.g., Watt-Meyer et al., 2021).

Watson (2019) examined the tendency error correction approach in the Lorenz 96 model. Brajard et al. (2021) explored the resolvent error correction approach in the two-scale Lorenz model as well as in a low-order coupled ocean-atmosphere model called the Modular Arbitrary-Order Ocean-Atmosphere Model~~(MAOOAM) (De Cruz et al., 2016)~~ (MAOOAM, De Cruz et al., 2016) . Watt-Meyer et al. (2021) demonstrated that the online error correction can improve the short-term forecasting ~~skills~~ skill and accuracy of precipitation simulation while the dynamical model can run indefinitely without numerical instabilities arising. Gregory et al. (2024) applied ML to correct sea ice errors in an ocean-ice coupled model and demonstrated that ML can effectively reduce sea ice bias ~~online~~ in a 5-year simulation. So far, the ML-based online error correction method has not been tested for seasonal sea ice prediction in an Earth system model. In this study, we ~~will~~ build and assess a hybrid model combining ML and a state-of-the-art Earth system model for seasonal prediction of Arctic sea ice.

On the other hand, the offline error correction consists in performing post-processing (also called calibration) of the dynamical model predictions (blue line in Figure 1). ML is trained to predict errors for time-averaged model outputs (e.g., daily or monthly outputs) and applied to correct errors present in raw predictions. The most common error correction methods employed in sea ice prediction (Bushuk et al., 2024) are relatively simple (e.g., correction of the mean error or a linear regression adjustment, Blanchard-Wrigglesworth et al., 2017). More recently, Palerme et al. (2024) applied ML to improve the skill of sea ice ~~concentration~~ forecasts on the weather timescale. Overall, they illustrated that ML-based offline calibration reduced the SIC prediction errors by 41% and the ice edge distance error by 44%. Their application is mainly focused on short-term sea ice prediction within 10 days in an ocean-ice coupled model. ~~We will~~ In this study, we apply and assess the ML-based calibration for seasonal prediction of Arctic sea ice in a state-of-the-art fully-coupled Earth system model.

In this study, we apply ML to the Norwegian Climate Prediction Model (NorCPM, Wang et al., 2019), a fully-coupled Earth system model, for seasonal prediction of Arctic sea ice. We test and compare the ML-based online and offline error correction approaches. In the online approach, we build a hybrid model combining ML and NorCPM to update the instantaneous sea ice state during the production of seasonal predictions. In the offline approach, we use ML to calibrate raw seasonal predictions of Arctic sea ice. The comparison between the two approaches within the same framework delivers new insights for the sea ice prediction community into how to effectively use ML for seasonal Arctic sea ice predictions.

The paper is organized as follows: section 2 presents the ~~model,~~ dynamical model, data, ML-based error correction approaches, experimental design, and metrics for validation. Section 3 shows the results of different experiments. We finish with ~~conclusions and discussions~~ discussions and conclusions in section 4.

2 ~~Methods~~ Data and ~~data~~Methods

2.1 Norwegian Climate Prediction Model

The dynamical model we used is NorCPM ~~(Counillon et al., 2014, 2016; Kimmritz et al., 2018, 2019; Wang et al., 2016, 2017)~~ (Counillon et al., 2014, 2016; Wang et al., 2016, 2017; Kimmritz et al., 2018, 2019). 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, Evensen, 2003)~~(EnKF, Sakov and Oke, 2008).

NorESM1 (Bentsen et al., 2013) is a fully-coupled Earth system model used for climate simulations. Its ocean component is the Bergen Layered Ocean Model (BLOM, Bentsen et al., 2013) – an updated version of the isopycnal coordinate ocean model MICOM (Bleck et al., 1995). The sea ice component is the Los Alamos sea ice model version 4 (CICE4, Gent et al., 2011; Holland et al., 2012). The atmospheric component is a variant of the Community Atmosphere Model version 4 (CAM4-Oslo, Kirkevåg et al., 2018). The land component is the Community Land Model ~~(CLM4, Lawrence et al., 2011; Thornton, 2010)~~ (CLM4, Thornton, 2010; Lawrence et al., 2011). Furthermore, the version 7 coupler (CPL7, Craig et al., 2012) is utilized for inter-component communication and interaction. The external forcings follow the protocol of the Coupled Model Intercomparison Project Phase 5 (CMIP5) historical experiment (Taylor et al., 2012).

The atmospheric and land components are situated on the National Center for Atmospheric Research (NCAR) finite-volume 2° grid, featuring a regular $1.9° \times 2.5°$ latitude–longitude resolution with 26 hybrid sigma–pressure levels extending to 3 hPa. The ocean and sea ice components utilize NCAR's gx1v6 horizontal grid, which is a nominal 2° resolution curvilinear grid with the northern pole singularity shifted over Greenland (Bethke et al., 2021). This grid is enhanced both meridionally towards the equator and zonally and meridionally towards the poles. The ocean component comprises 51 isopycnic layers, featuring a bulk mixed layer representation on top with two layers having time-evolving thicknesses and densities.

The sea ice component is equipped with ~~several~~ five ice thickness categories ~~(we use the predefined value of N=5)~~ to account for the different thermodynamic and dynamic properties of ice with different thicknesses. The volume of snow and ice, energy content, as well as ~~ice concentration~~SIC, surface temperature, and the volume-weighted mean ice age are determined for each of the ice thickness categories (Bentsen et al., 2013; Kimmritz et al., 2018, 2019).

NorESM1 tends to overly produce thick sea ice, especially in the polar oceans adjacent to the Eurasian continent. This is partly due to factors such as weaker winds across the polar basin and overestimated Arctic cloudiness, which ~~slows~~ leads to little summer snowmelt. Consequently, the summer ~~sea ice extent~~ SIE in the Arctic ~~is too large~~ has large positive biases, contributing to an underestimation of global temperatures (Bentsen et al., 2013; Bethke et al., 2021).

NorCPM uses ~~an EnKF-based anomaly-field DA scheme~~ the EnKF to update unobserved ocean and sea ice variables by leveraging state-dependent covariance from the simulation ensemble (Kimmritz et al., 2018, 2019). The EnKF allows the assimilation of observations of various types while accounting for observational errors, spatial coverage, and the evolving covariance with the climate state. The EnKF ~~ensures accurate ensemble predictions by representing uncertainties in the initial conditions , propagating these uncertainties over time , and providing a spatiotemporal estimate~~accounts for uncertainties in initial conditions to generate ensemble predictions, which evolve in time and provide time- and space-dependent error estimates.

NorCPM employs anomaly-field assimilation (Kimmritz et al., 2019; Wang et al., 2019; Bethke et al., 2021) in which the climatology of the observations is replaced by the model climatology calculated from the ensemble mean of the model historical simulation (without assimilation). While the anomaly-field assimilation keeps the model close to its attractor and helps to reduce the model drift during the monthly model integration (Carrassi et al., 2014; Weber et al., 2015), it does not significantly change model biases.

2.2 Data

~~The reanalysis of NorCPM combining observations with NorESM is a physically consistent construction of the Earth system (Counillon et al., 2016; Kimmritz et al., 2019) and represents the upper limit of the sea ice predictability of NorCPM.~~ In this study, we use the reanalysis of NorCPM as the "truth" to assess the improvement achieved by the ~~error correction methods.~~ ML-based error correction approaches. First, it is because NorCPM performs anomaly-field assimilation. The large model biases are not corrected by DA (section 2.1) and thus the analysis increment of the reanalysis used to build the online error correction model (section 2.3) does not take into account model biases. Second, the online error correction approach needs to consistently update SIC in each category, sea surface temperature (SST), and sea surface salinity (SSS) under sea ice, which are often not observed. The reanalysis of NorCPM is a physically consistent construction of the Earth system (Counillon et al., 2016; Kimmritz et al., 2019) and provides a reasonable and physically consistent estimation of these variables. Finally, the reanalysis combining observations with NorESM represents the upper limit of the sea ice predictability of NorCPM.

The reanalysis is available from 1980 to 2021 with 30 ensemble members. The initial states of the ~~reanalyses~~ 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 ~~sea surface temperature (SST)~~SST, SIC, and subsurface hydrographic profile data in the middle of each month.

~~The DA setting slightly differs for two different periods 1980-2002 and 2003-2021.~~ From 1980 to 2002, the climatology used for ~~anomaly~~ 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 ~~data assimilation occurs, updating~~ DA occur and update the sea ice and ocean states within the ocean mixed layer. From 2003 to 2021, the climatology utilized for ~~anomaly~~ 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 ~~in a single step~~ 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.

2.3 Online error correction approach

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).

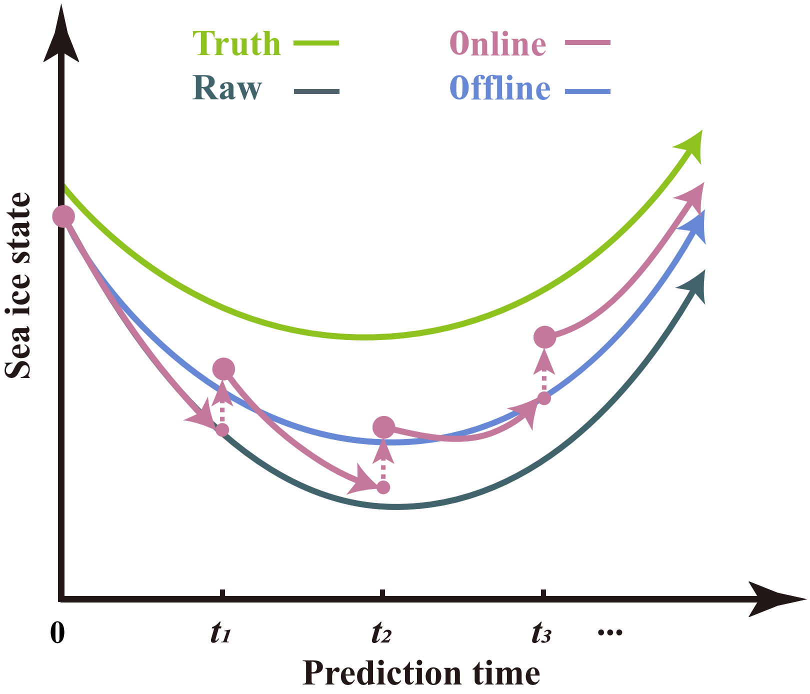

Figure 1. Schema for the online and offline ML-based error correction ~~methods~~approaches. The ~~pink~~ green line represents the "truth". The gray line represents dynamical prediction without error correction. The purple (blue) line represents prediction with online (offline) ML-based error correction. The purple dashed arrows indicate pauses during the prediction production, facilitating correction to the instantaneous model ~~state~~states.

The ~~reanalysis (described in~~ monthly model integration of the reanalysis (section 2.2) ~~is used to produce a forecast following:~~ can be described as follows:

$$\mathbf{x}_k^f = \mathcal{M}(\mathbf{x}_{k-1}^a), \qquad (1)$$

where $\mathbf{x}_k^f$ represents the forecasted instantaneous model state at t_k, $\mathcal{M}$ represents the dynamical model integration from time t_{k-1} to t_k (section 2.1). During the analysis, DA uses available observations to generate $\mathbf{x}_k^a$ — an updated instantaneous model state and initial conditions for the next monthly model integration from time ~~t_{k-1} to time t_k~~t_k to time t_{k+1}.

The online approach is to emulate the ~~analysis increments of DA defined as $\mathbf{x}_k^a - \mathbf{x}_k^f$~~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 ~~correction~~ prediction model can be expressed as~~follows:~~:

$$\varepsilon = \mathcal{M}_{\mathrm{e}}(\mathbf{x}^f), \qquad (2)$$

where $\mathcal{M}_{\mathrm{e}}$ represents the data-driven model taking the instantaneous model state $\mathbf{x}^f$ as input and ε represents the ~~estimated~~ predicted model error.

The hybrid model, incorporating the dynamic model and the online error correction model, can be expressed as follows:

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

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. Considering the seasonality of the error of the sea ice state, we build one error ~~model (Eq. 2)~~ correction model for each calendar month. Also, we employ a running training strategy and use the most recent 11 years of data before the prediction month (the first 10 years for training and the last year for validation). The input feature contains latitude, SST, SSS, five categories of SIC, and five categories of sea ice volume in the middle of the month. The output feature consists of errors in SST, SSS, and 5 categories of SIC (Table 1). Please refer to section 2.5 for the ML configuration.

~~The hybrid model incorporating the dynamic model and the~~ Before restarting the model after applying online error correction~~model can be expressed as follows:~~

~~$\mathbf{x}_l^h = \mathcal{M}(\mathbf{x}_{l-1}^h) + \mathcal{M}_\mathrm{e}(\mathbf{x}_l^f)$~~

~~where $\mathbf{x}_l^h$ represents the error-corrected instantaneous model state at t_l during the prediction.~~

~~It is crucial to ensure the physical consistency between the ice and ocean components. Inconsistencies can arise after online error correction, such as through the analysis variablesupdate, which may yield unphysical values , or when certain variables remain uncorrected and must be diagnosed from the updated variables. The~~ , 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 ~~is~~ specifically designed for NorCPM ~~(Kimmritz et al., 2018)~~(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 constraint and model stability after the error correction.

2.4 Offline error correction approach

The offline error correction approach refers to performing post-processing of the dynamical model predictions (blue line in Figure 1). The ML configuration is the same as the online configuration (section 2.5). 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. For more details about the offline error correction approach, please refer to Table 1.

Table 1. Information about online and offline ML-based error correction models

	Online ML-based model	Offline ML-based model
Input features	Instantaneous SST, SSS, latitude, 5 categories SIC, and sea ice volume	Monthly SST, SSS, latitude, SIC, and sea ice volume
Output features	Instantaneous SST, SSS, and 5 categories SIC errors	Monthly SIC prediction error
Data	The most recent eleven years data (ten years for training and one year for validation)	
Remark	Only apply to sea-ice covered grids in the Arctic with SIC values greater than 1%.	

It's worth noting that the offline error correction approach targets directly monthly average model outputs, whereas the online error correction approach addresses instantaneous model errors (Figure 1) and indirectly changes the monthly model outputs during the production of predictions. Therefore, their input and output features are different (Table 1).

2.5 Machine learning configuration

As mentioned in the previous sections, the ML model configurations employed for online and offline error correction approaches ~~differ in the input and output variables (for more details, please refer to Table 1), but~~ share an identical architecture (i.e., the same number of layers and the same number of neurons in each layer), but differ in the input and output variables, resulting in different numbers of trainable parameters (for more details, please refer to Table 1).

The ML model uses the values from a single grid point as input to predict the value at the same grid point, meaning one ML model for all grid points. This simplifies the training process while still enabling the development of efficient models.

The ML architecture ~~employed~~ used in this study is a multilayer perceptron (MLP), a ~~powerful model known for its ability to capture complex nonlinear relationships in data. As a~~ fully connected neural network ~~, MLP excels in function approximation, making it particularly~~ well-suited for ~~error correction in geophysical modeling (Yang et al., 2023). Its key advantagesinclude~~ capturing complex nonlinear relationships in data. MLP offers several advantages, including flexibility in handling diverse input features, efficient training ~~through~~ via backpropagation, and strong generalization when properly regularized. ~~Moreover~~Additionally, MLP is computationally ~~efficient compared to more~~ more efficient than complex deep learning architectures ~~. As noted by Jia et al. (2019) and Watson (2019), error-correcting learning problems generally require smaller ML models and fewer training data , making MLP a practical choice for integrating data-driven corrections into NorCPM~~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) .

The entire MLP architecture consists of ~~five~~ seven layers:

Table 2. 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

- ~~The input layer includes a~~ **Input layer**: A batch normalization layer (Ioffe, 2017), which helps ~~to regularize and normalize~~ stabilize and accelerate the training process by normalizing the input features.
- ~~The second layer is a~~ **Second layer**: A dense layer with 60 neurons~~. It applies~~, using the rectified linear unit (ReLU) activation function~~, which introduces non-linearity into the network.~~.
- ~~The third layer has the same configuration~~ **Third layer**: A dense layer with 30 neurons, also employing the ReLU activation function. This layer shares the same structure as the second layer~~, with~~.
- **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~~function~~, identical to the third layer.
- ~~The fourth layer is the attention layer , which is used for helping better training.~~ **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:

- ~~The output layer is a dense layer with a linear activation function.~~ **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.

~~We~~ To achieve better training results, we further implement the following settings:

- We ~~employ~~ adopt a running training ~~set approach~~ strategy, using data from the ~~most recent~~ 11 years ~~. For example, to build an error correction model for~~ 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 ~~is used for training , and data in~~ 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 ~~decided to only use the data close to the test period. We also performed a sensitivity study~~ also performed sensitivity studies on the length of the running training set (e.g., the most recent 5 years or all years since 1980) ~~, which is~~ 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 ~~in~~ from the most recent 11 years leads to the best performance for ML training, and the running training outperforms the fixed-period training.

- The characteristic of model errors varies with the calendar month. For instance, the model errors mainly appear in the marginal zone in winter but in the entire sea ice-covered region in summer. We train separately for each calendar month, leading to a distinct ML model for each calendar month. This results in 236 neural network models (from February 2003 to September 2022 based on test months) for the online case. In the offline case, we ~~consider also~~ also consider the start month, resulting in 836 (4 initialized months × 11 lead months × 19 test years) models. Despite the large number of models, the training process is highly efficient due to the simple architecture and low data dimensionality. As a result, training each model is very quick, taking only one minute on a CPU, making this exhaustive approach computationally affordable.

- We train and apply ~~an error correction model~~ error correction models to grid points where the total SIC exceeds 1%. It avoids adding sea ice into open water areas and thus dynamical inconsistency. It also means that our correction ~~model~~ models can not create ice on a grid point where the model predicted ice-free conditions.

2.6 Hindcast experiments

The standard hindcasts (hereafter referred to as **Reference**) are initialized from the reanalysis presented in section 2.2 in the middle of January, April, July, and October each year, spanning from ~~1991~~ 1992 to 2021, with a duration of 12 months. From ~~1991~~ 1992 to 2002, the first 9 ensemble members of the 30-member reanalysis are used to carry out the hindcast experiments, while after 2003, ~~its~~ the first 10 ensemble members are used to initialize the hindcast experiments. It is worth noting that these differences (i.e.~~the number of ensemble members~~, the different ensemble sizes) would have minimal impact on the results of this study.

A new set of hindcasts (hereafter referred to as **OnlineML**), similar to Reference but with the online error correction approach (section 2.3), are initialized from the reanalysis in the middle of January, April, July, and October from 2003 to 2021. In the production of ~~a~~ each hindcast, NorCPM pauses in the middle of each lead month and uses the online error correction model ~~(Eq. 2)~~ to predict the error correction and then update the instantaneous model state.

The offline error correction approach (section 2.4) is applied to post-process the hindcasts of Reference (hereafter referred to as **OfflineML**).

2.7 Metrics for evaluation

SIE is a commonly used metric in seasonal sea ice prediction (e.g., Bushuk et al., 2024). We evaluate the prediction skill of SIE in the ~~Pan-arctic and in~~ Pan-Arctic and six Arctic regions depicted in ~~Fig.~~ Figure 2. These regional ~~delineations~~ definitions adhere to the area definitions provided by Kimmritz et al. (2019), albeit with the consolidation of the original fourteen sea areas into six regions that are very similar to the ones used in Bushuk et al. (2024). In this study, ~~we compute an areal sum~~ the SIE is defined as the total area of all grid points ~~in~~ within the region of interest ~~with SIC ≥~~ where SIC ≥ 15%. SIE is calculated for each ensemble member~~and evaluate the average of SIEs over different ensemble~~ , and we evaluate the ensemble mean by averaging SIE across all ensemble members.

To evaluate the ~~prediction skill of SIE~~performance of the ML-based error prediction models, we employ the mean absolute error (MAE), defined as:

$$\mathrm{MAE} = \frac{1}{M}\sum_{i=1}^{M}|\mathbf{E}_{\mathrm{p}} - \mathbf{E}_{\mathrm{t}}|, \tag{4}$$

where $\mathbf{E}_{\mathrm{p}}$ denotes the predicted error and $\mathbf{E}_{\mathrm{t}}$ denotes the "true" error. In more detail, $\mathbf{E}$ refers to the SIC error at each grid point over the entire evaluation period. M represents the total number of data points used in the MAE calculation.

To evaluate the sea ice prediction skill, we employ the root mean square error (RMSE) as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathbf{X}_{\mathrm{prediction}} - \mathbf{X}_{\mathrm{reanalysis}})^2}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathbf{X}_{\mathrm{p}} - \mathbf{X}_{\mathrm{t}})^2}, \tag{5}$$

where ~~$\mathbf{X}_{\mathrm{prediction}}$~~ $\mathbf{X}_{\mathrm{p}}$ represents the prediction and ~~$\mathbf{X}_{\mathrm{reanalysis}}$ represents the reanalysis~~ $\mathbf{X}_{\mathrm{t}}$ represents the "truth" (i.e., the ~~truth~~ reanalysis in this study). In this ~~context~~study, $\mathbf{X}$ ~~could~~ can refer to either the ~~SIE in the~~ integrated ice-edge error (IIEE)

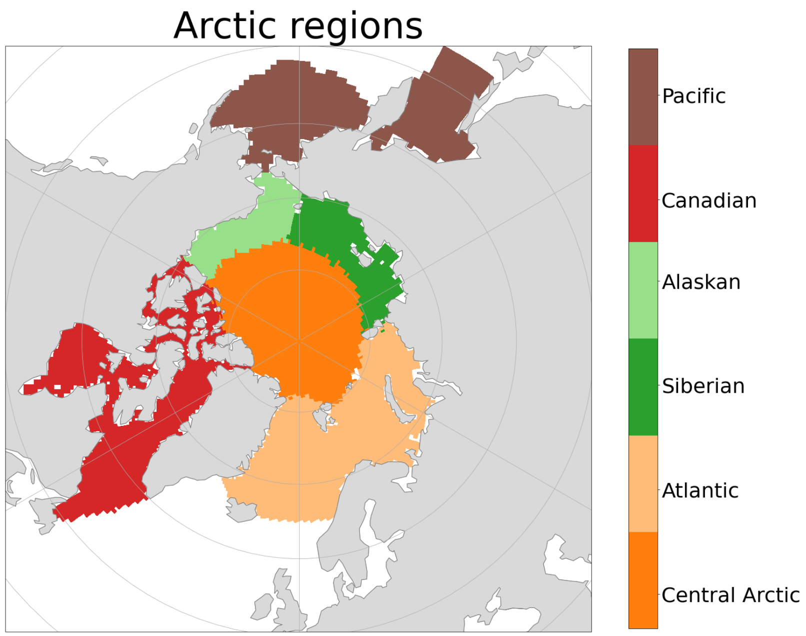

Figure 2. Regional domain definitions for ~~Central~~ central Arctic, Atlantic, Siberian, Alaskan, Canadian, and ~~Regions~~ Pacific regions are based on sea area definitions in Kimmritz et al. (2019) and are similar to those used in Bushuk et al. (2024). Atlantic region: ~~GIN Sea~~Greenland, ~~Barents Sea~~Ice, Norwegian, Barents and Kara ~~Sea~~Seas; Siberian region: Laptev ~~Sea,~~ and East Siberian ~~Sea~~Seas; Alaskan region: Chukchi ~~Sea,~~ and Beaufort ~~Sea~~Seas; Canadian region: Canadian archipelago, Hudson Bay, Baffin Bay, and Labrador Sea; Pacific region: Bering Sea ~~,~~ and the Sea of ~~the~~ Okhotsk.

on a Pan-Arctic scale, the SIE on a Pan-Arctic/regional scale, or the SIC at ~~each~~ a specific grid point. N ~~denotes the length of the experiment period, which spans~~ represents the number of hindcasts, spanning from 2003 to 2021.

The ~~integrated ice-edge error (IIEE)~~ IIEE is also a crucial metric for sea ice predictions (Goessling et al., 2016). It specifically captures the discrepancies along the ice edge by quantifying the area where the predicted and ~~true ice concentrations~~ "true" SIC differ significantly. This makes IIEE particularly valuable for evaluating the spatial accuracy of the ice edge location, offering insight into the performance of models in reproducing the dynamic boundary between ice-covered and open ocean regions. ~~We define the IIEE~~ Following the definition of Goessling et al. (2016), the IIEE is computed as the area where the prediction and the ~~truth~~ "truth" disagree on the ~~ice concentration~~ SIC being above or below 15%:

$$\mathrm{IIEE} = \int_A \max(c_\mathrm{p} - c_\mathrm{t}, 0)\, dA + \int_A \max(c_\mathrm{t} - c_\mathrm{p}, 0)\, dA, \tag{6}$$

where A is the area of grid cell, $c = 1$ where ~~the sea ice concentration is above 15%~~ SIC is above 15% and $c = 0$ elsewhere, and subscripts p and t denote the prediction and the ~~truth~~"truth". The definition of the IIEE is equivalent to the so-called symmetric difference between the areas enclosed by the predicted and ~~true~~ "true" ice edges.

To evaluate the significance of prediction skill difference, we use a two-tailed Student's t-test to compare the IIEE or the ~~squared errors of SIE~~ RMSE between two predictions. ~~This statistical test determines whether the difference between the two sets is statistically significant.~~

To estimate the uncertainties ~~of the RMSE , we apply~~ in an RMSE value arising from the small ensemble size, we employ the bootstrap method. Specifically, we randomly ~~select, with replacement ,~~ sample 10 ~~data points from the 10 ensemblemembers, calculate~~ ensemble members with replacement from the ensemble, compute the ensemble mean, and ~~compute~~ then calculate the RMSE (for either SIC or SIE) ~~for this sampled~~ 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 ~~values~~value.

3 Results

3.1 Error correction model performance

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

The ~~‘true’~~ "true" errors obtained from analysis increments and the errors predicted by the online error correction model are averaged over 2003-2021 and displayed in ~~Fig.~~Figure 3. The ~~error patterns~~ spatial patterns of the "true" error vary significantly across different dates. For instance, on August 15, ~~the~~ errors are predominantly ~~positive in most regions due to that~~ negative across most regions as NorCPM underestimates SIC in this month, with some ~~negative~~ localized positive errors occurring internally. The average MAE across ice-covered grid points is 0.24%. On October 15, the errors ~~shift to negative, due to SICoverestimation in NorCPM~~are mostly positive as NorCPM overestimates SIC, resulting in an average MAE of 0.22%. On December 15, the ~~errors approach zero, appearing mainly in the marginal~~ MAE is 0.20%, primarily appearing in marginal ice areas, with ~~most grid points showing lower error~~ overall lower magnitudes compared to August and October. ~~Additionally, the errors along the ice edge remain consistently positive~~Notably, the average error remains below 1% in all cases.

For all those months and regions, ~~it shows that~~ the online error correction models can correctly predict the spatial pattern of the ~~prediction~~ "true" error (Figure 3~~d-f). The~~). Also, the magnitude of the "true" error is well reproduced with a slight underestimation.

To assess the offline error correction model, we show its performance for hindcasts initialized in July (Figure 4). ~~Similar to the instantaneous fields (i.e., online errors in Figure 3), the monthly~~ The monthly "true" error patterns vary significantly across months. The offline error correction ~~model effectively predicts~~ models effectively predict the spatial pattern of the ~~prediction~~ "true" errors (Figure 4~~d-f~~). The ~~prediction~~ predicted error magnitude is ~~also well captured~~similar to that of the "true" error, with only a slight underestimation. Notably, the MAE of the offline error correction approach is higher than that of the online error correction approach in December (0.30% versus 0.20%), which can be attributed to the model divergence since the initialization in July.

In summary, the above results suggest that the ~~ML~~ ML-based error correction models in both online and offline scenarios can skillfully predict the large-scale spatial patterns of the SIC error, but slightly underestimate its magnitude.

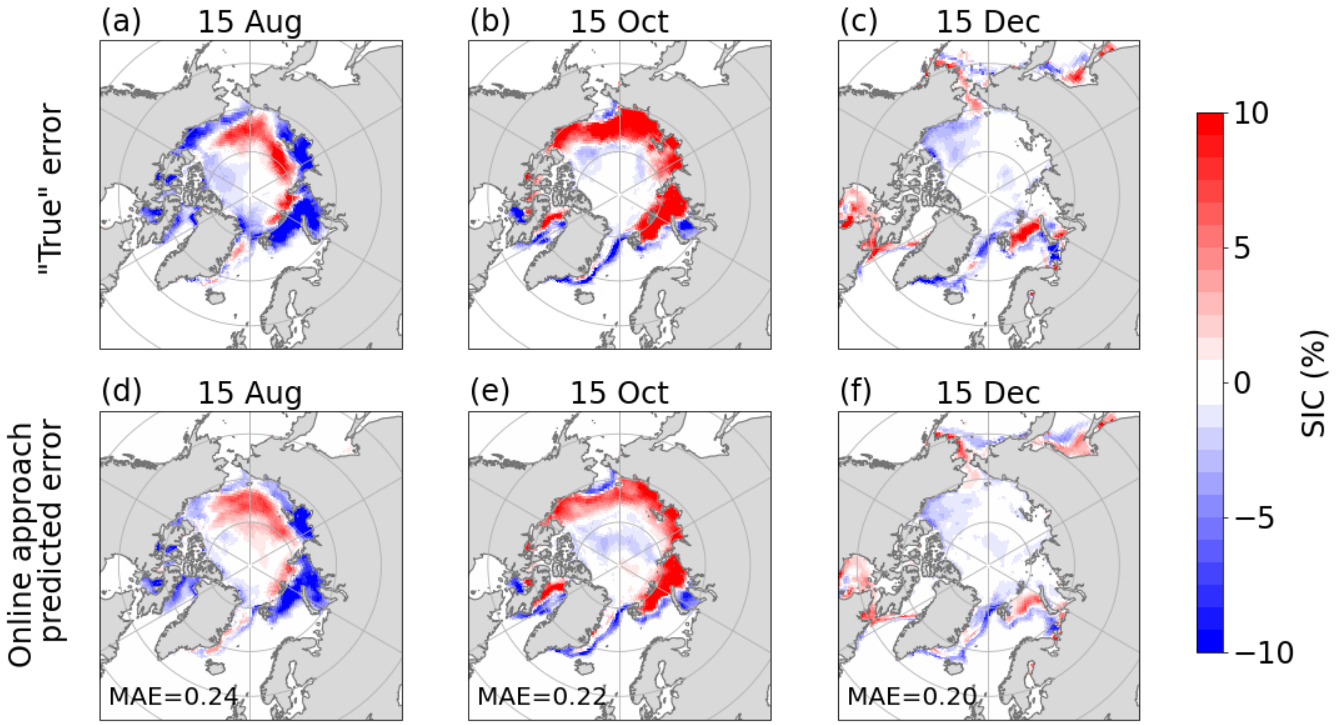

Figure 3. Top row: "true" errors of SIC in the middle of the month based on the analysis increments (i.e., the changes thanks to monthly DA in the reanalysis). Bottom row: the errors predicted by the online error correction model. These errors are averaged over the period ~~2003–2021~~2003-2021. Values in the bottom row are the MAE between the "true" and predicted errors across space.

3.2 Application ~~into~~ to seasonal predictions

3.2.1 Skill seasonality

~~In this section , we assess~~ This section assesses the three sets of hindcasts initialized in January, April, July, and October from 2003 to 2021. The ensemble hindcasts are initialized with the first 10 members of the reanalyses and predict for 11 months (section 2.6).

Figure 5 presents a comparative analysis of the RMSE for SIE prediction and the IIEE for ice edge prediction in the ~~pan-Arctic~~ Pan-Arctic across the three hindcast sets. The Reference hindcast shows higher RMSE in September and October (Figure 5a), primarily due to several factors that have been documented in Bentsen et al. (2013). NorCPM overestimates the Arctic cloudiness, and its summer-season snowmelt is too slow. In addition, NorCPM has slightly too weak winds across the polar basin. These factors lead to too thick sea ice in the polar oceans and excessive Arctic SIE, in particular in summer (Bentsen et al., 2013).

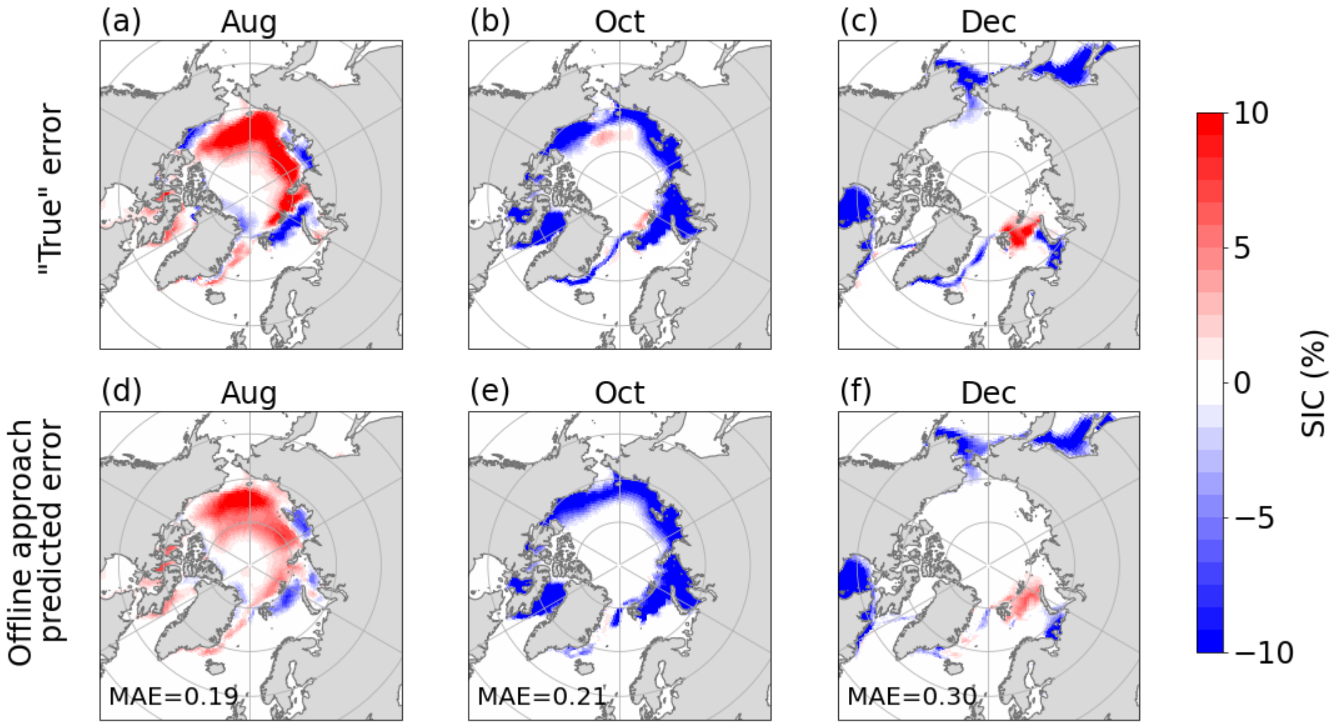

Figure 4. Top row: "true" errors of monthly SIC estimated by the ~~reanalysis minus the reference~~ Reference hindcast initialized in July minus the reanalysis. Bottom row: the errors predicted by the offline approach. The errors are averaged over the period ~~2003–2021~~2003-2021. Values in the bottom row are the MAE between the "true" and predicted errors across space.

Both the OnlineML and OfflineML hindcasts exhibit ~~similar behaviors regardless of the seasonality (Figure 5b and 5c):~~ a small error reduction from January to July and a large error reduction from August to December ~~. For OnlineML hindcasts, although only the errors in Arctic~~ (Figure 5b and 5c). The OnlineML hindcast, in which only SIC, SST, and SSS are corrected without ~~adjusting atmospheric model errors, the predictions show~~ directly adjusting the atmospheric component, shows some improvements, particularly in January and from September to December. In contrast, from February to August, the Reference hindcast already exhibits good performance, leading to no significant differences. Compared ~~with the OnlineML hindcasts~~to the OnlineML hindcast, the OfflineML ~~hindcasts have a larger~~ 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. The primary reason is that the online approach corrects instantaneous model errors (on the 15th day of the month). Still, during the one-month model integration, the sea ice component dynamically interacts with the other components, leading to error growth. In terms of monthly averaged model outputs, the correction ~~is likely~~ magnitude is damped. In contrast, the offline approach aims to directly post-process monthly outputs without model integration.

The IIEE shows similar results to the RMSE of SIE (Figure 5d-f). For the Reference hindcast, the IIEE is higher from July to September. The online approach leads to some improvements over the Reference hindcast from July to December,

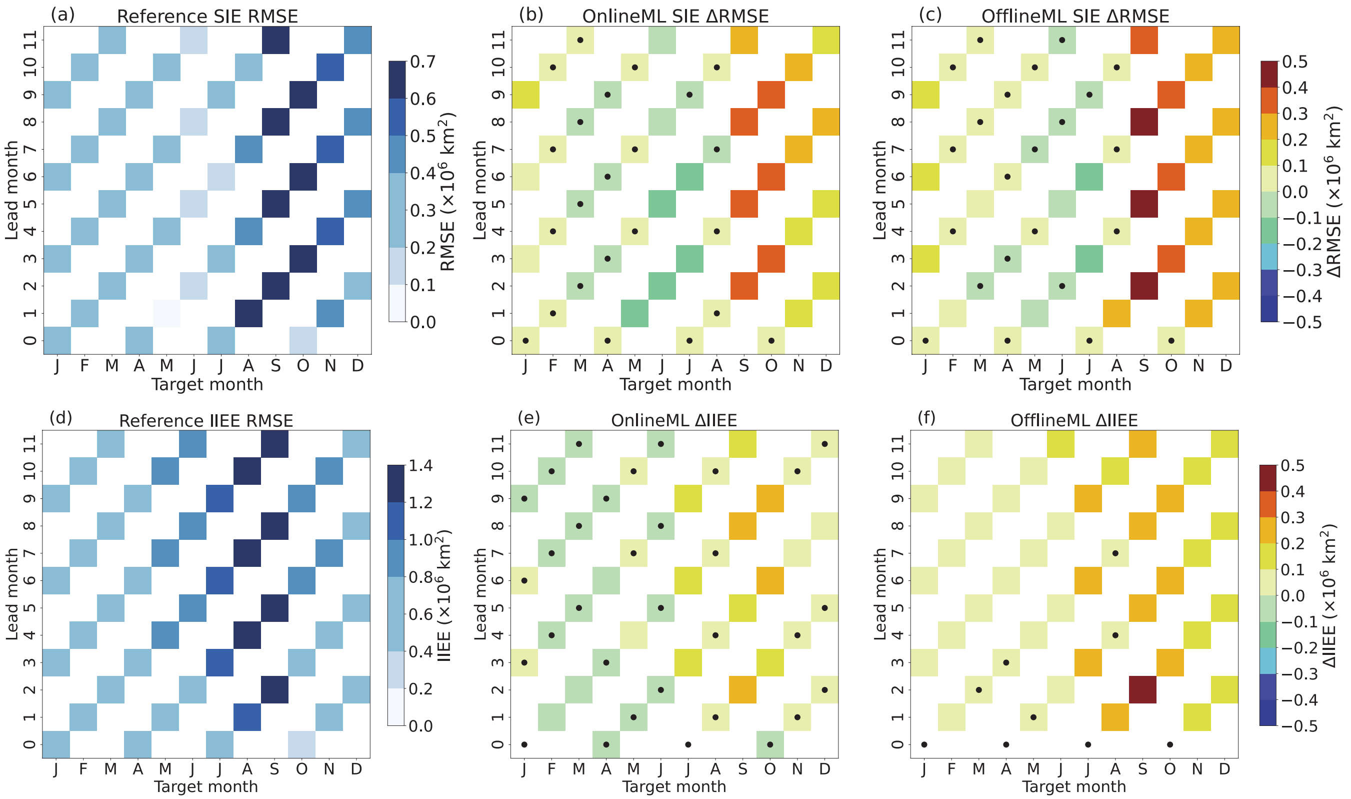

Figure 5. (a) RMSE of SIE for the Reference hindcast, (b) ΔRMSE between the Reference and OnlineML hindcasts, (c) ΔRMSE between the Reference and OfflineML hindcasts. (d) IIEE of the Reference hindcast, (e) ΔIIEE between the Reference and OnlineML hindcasts, (f) ΔIIEE between the Reference and OfflineML hindcasts. In b,c,e, and f, warm colors (red/yellow) indicate that the OnlineML or OfflineML hindcasts are better than the Reference hindcasts, while cold colors (blue/green) indicate they are worse than the Reference hindcast. The black dots represent regions where the ΔRMSE or ΔIIEE does not pass the 95% significance test.

but its error reduction is small or not significant in the other months. In contrast, the offline approach consistently improves ~~the~~ performance across nearly all periods and ~~demonstrates~~ 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 ~~shows~~ exhibits relatively larger prediction errors from August to October, primarily due to increased model uncertainties ~~related to atmospheric~~ associated with atmospheric forcing and sea ice processes. The offline approach outperforms the online approach in reducing both ~~RMSE for SIE and IIEE for~~ the RMSE of SIE and the IIEE along the ice edge, ~~especially in monthswith higher prediction errors.~~ 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.

3.2.2 Skill of seasonal predictions for different regions

The previous section highlighted significant improvements in predictions, primarily evident from September to January regardless of the initialized month. In this section, we focus on analyzing the hindcasts initialized in July, and we show the performance for different regions and both SIE and SIC. It is mostly because summer sea ice prediction serves as a critical climate change indicator, affects ecosystems and human activities, and presents a significant scientific challenge due to its high variability (Figure 5a). For validation on the other initialization months, please refer to Figures S1-S4.

We first investigate the seasonal prediction skill for ~~pan-Arctic~~ Pan-Arctic and regional SIE defined in Figure 2. For the ~~pan-Arctic~~ Pan-Arctic SIE, previously assessed in ~~Fig.~~ Figure 5, both the OnlineML and OfflineML hindcasts reduce the SIE RMSEs (Figure 6a). The RMSEs in the OnlineML hindcast have a strong seasonality as that in the Reference hindcast: higher in August, September, and October, and lower in November, December, and January. The OfflineML hindcast has the lowest RMSEs, in particular, an RMSE reduction of about ~~70~~75% compared to the Reference ~~hindcasts~~ hindcast in September.

Both error correction approaches reduce the RMSEs for regional SIE, and the offline approach overall outperforms the online approach (Figure 6b-f). In the Atlantic region (Figure 6b), significant RMSE reduction is observed for the ~~three first~~ first three months, until October. The OfflineML hindcast has the lowest RMSEs until September and similar RMSEs to the OnlineML hindcast from October. In the Siberian region (Figure 6c), the RMSE reduction due to error correction is significant only until ~~Ocotber~~ October but becomes almost ~~zeros~~ zero from November due to the region being fully covered by sea ice. The OfflineML hindcast is significantly better than the OnlineML hindcast until September and similar ~~aftetwards~~afterward. In the Alaskan region (Figure 6d), there is no significant RMSE reduction in August, but we observe significant RMSE reductions from September to November. In December and January, the region is almost fully covered by sea ice, leading to very tiny RMSEs for all three hindcast experiments. In the Canadian region (Figure 6e), both approaches lead to significant RMSE reductions from September to December and the offline approach outperforms the online approach. In addition, the online approach leads to a significantly larger RMSE in August than that of the Reference hindcast. In the Pacific region, the RMSEs are near zeros from August to October due to very limited sea ice coverage. The two error correction approaches lead to significant RMSE reductions after November, and the offline approach outperforms the online approach in December and January.

~~Compared with the online approach, the offline approach performs better in each region. The primary reason is that the online approach corrects~~ Notably, in August, the RMSE of the OnlineML hindcast exceeds that of the Reference hindcast in both the Alaskan and Canadian regions. This is primarily due to the systematic underestimation of SIE by the OnlineML hindcast relative to both the Reference hindcast and the "truth" in these regions (Figure S5). The underlying causes of this systematic underestimation, however, warrant further investigation.

The offline approach outperforms the online approach across all regions, primarily because the online correction targets instantaneous model errors (i.e., those on the 15th day of ~~the~~ each month). ~~The~~ These corrected errors may reemerge ~~due to errors from other components , damping the~~ through interactions with the other components of the coupled model system, thereby diminishing the overall impact of the online error correction when ~~computing monthly averaged model~~ evaluated

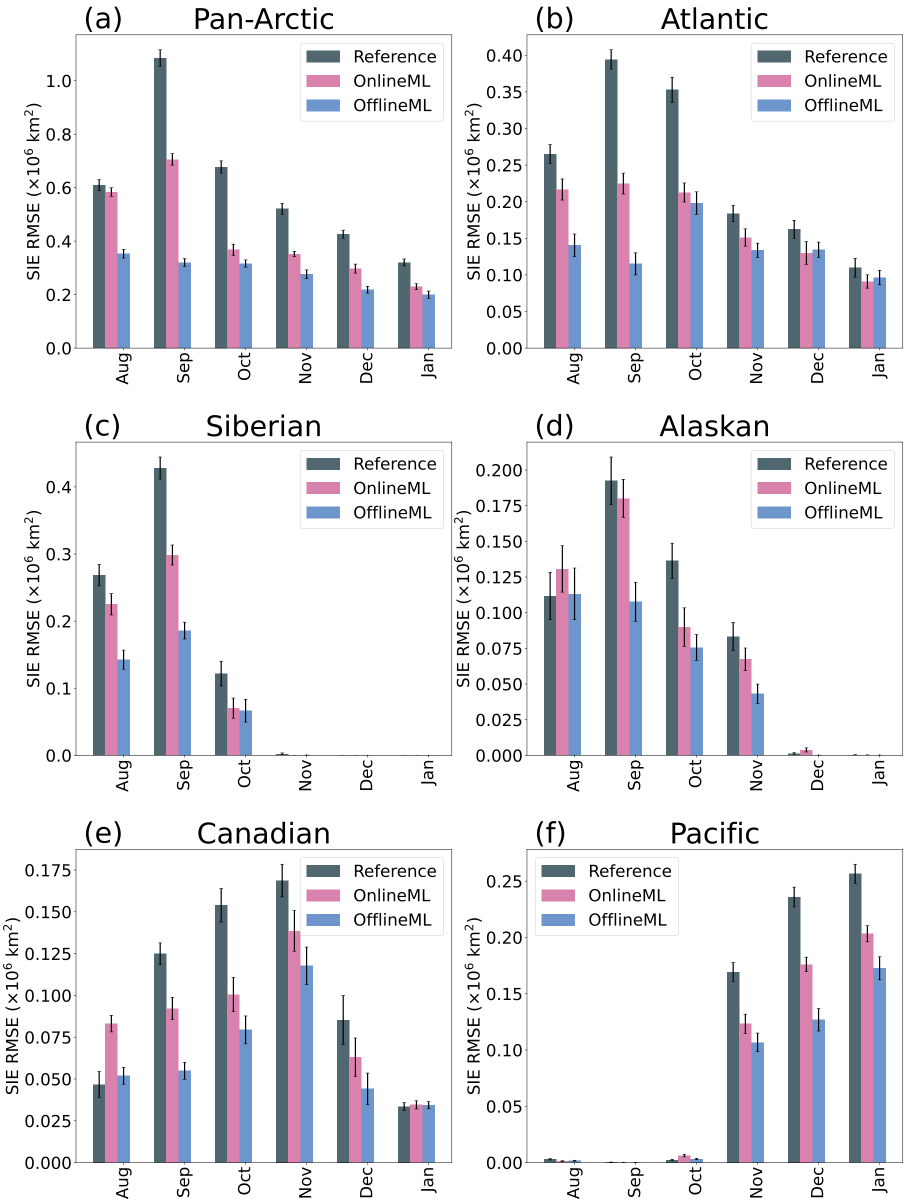

Figure 6. RMSE of SIE in the Pan-Arctic and five subregions for the Reference hindcast (gray bar), the OfflineML hindcast (blue bar), and the OnlineML hindcast (purple bar). Error bars represent the uncertainties.

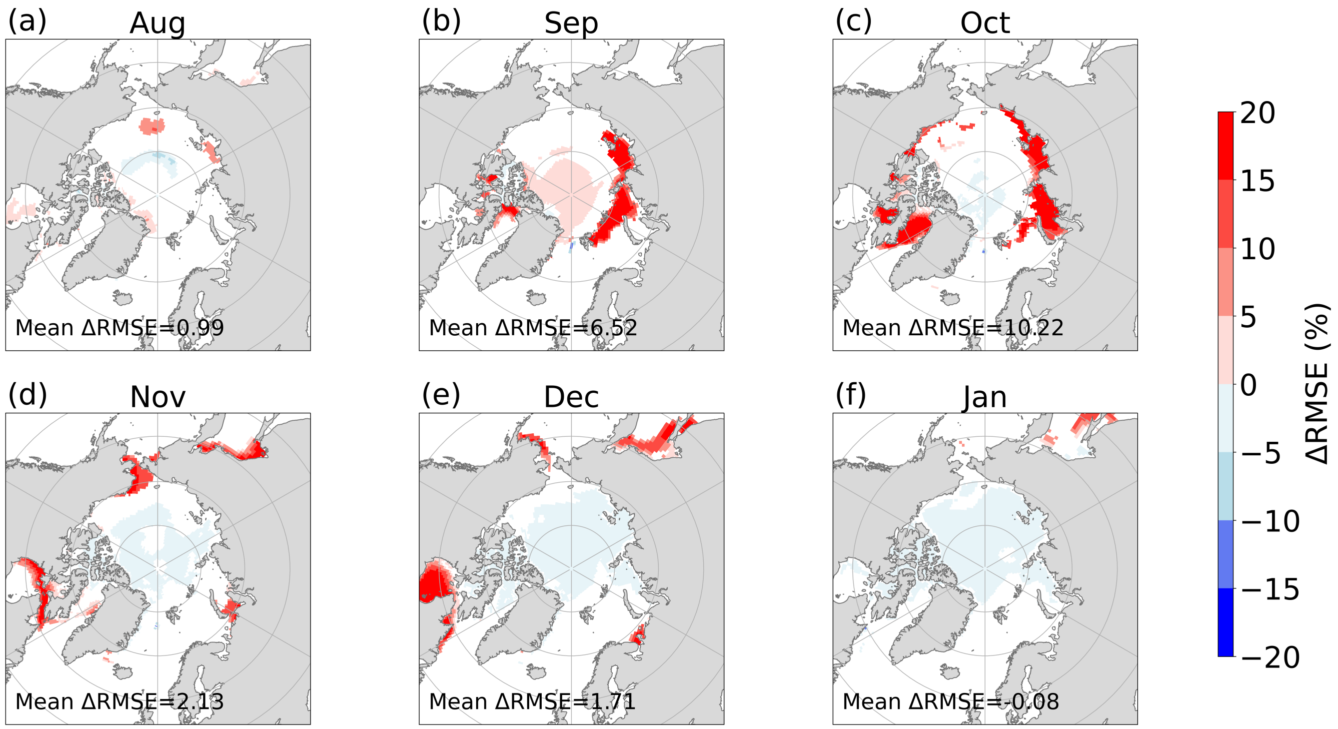

Figure 7. Differences ~~between~~ in SIC RMSE ~~of~~ between the Reference and OfflineML hindcasts initialized from July. Warmer (colder) colors indicate that the OfflineML hindcast ~~performs better~~ outperforms (~~worse~~underperforms) the Reference hindcast. White areas indicate differences that are not statistically significant. The ~~white color indicates~~ value in each subplot represents the ~~differences don't exceed the~~ average ΔRMSE for statistically significant ~~test~~grid points.

using monthly-averaged outputs. In contrast, the offline approach directly ~~corrects~~ adjusts the monthly model outputs, which ~~is consistent with the findings for the whole Arctic described in the previous section.~~ aligns closely with the evaluation metrics used in this study. Moreover, the offline approach does not need to run the dynamical model and is computationally cheaper than the online approach. However, the online approach not only reduces SIC errors but also propagates corrections through the model integration to the other variables (e.g., sea ice thickness and sea ice drift), ensuring physical consistency between the predicted variables.

In summary, while the error correction performance varies by region and target month, overall, ~~it improves the prediction of SIE~~both approaches improve the sea ice prediction. In addition, the offline approach is more efficient than the online approach in reducing the SIE ~~RMSEs for both pan-Arctic~~ RMSE for both Pan-Arctic and subregions. ~~For the results of the~~ These conclusions also hold for seasonal predictions initialized in the other seasons. For details, please refer to Figures S1-S4.

We take a closer look at the spatial aspects of the offline error correction approach in hindcasts initialized in July (Figure 7). We specifically focus on identifying local areas where the error correction leads to improvements that may not be evident when examining SIE alone.

The ~~improvements in SIC due to error correction are more discernible~~ 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 ~~region and the Atlantic region showed~~ and Atlantic regions exhibit improvements (Figure 7a), with an average improvement of 0.99%. In September, ~~significant improvements are observed, particularly in~~ notable enhancements appear across the central Arctic, Atlantic, Siberian, and Canadian regions (Figure 7b), with an average improvement of 6.52%. In October, significant ~~enhancements~~ improvements are observed in the Atlantic and Canadian ~~areas~~regions, reaching an average of 10.22%. Additionally, some ~~gray areas appear~~ blue areas emerge in the central Arctic, indicating ~~significant~~ substantial differences between the Reference hindcast and the OfflineML hindcast, though the magnitude of these differences ~~is very small. This phenomenon does not occur in the OnlineML hindcast (Figure S5). It suggests that, due to the lack of dynamical consistency, OfflineML introduces noises in fully ice-covered regions, but this noise is small.~~ remains minimal. In November and December, the positive impact of the error correction is ~~concentrated in areas like the~~ primarily concentrated in Hudson Bay and the ~~Okhotsk Sea . In January, the improvements only appear at a few locations within the Okhotsk Sea.~~ 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.

The major improvements in SIC are evident near the ice edge, which is closely associated with SIE. This spatial distribution highlights how ~~error correction~~ the ML-based error correction approaches can enhance model performance in different regions, particularly during the ice-advance season, where ~~broader SIE metrics~~ using SIE as a metric might obscure these localized gains. ~~It~~ In addition, it is noteworthy that the OfflineML and OnlineML hindcasts exhibit similar error spatial distributions. For specific details about the OnlineML hindcast, please refer to Figure ~~S5~~S6.

4 ~~Discussion~~ Discussions and ~~conclusion~~conclusions

In this study, we apply ML ~~in NorCPM–~~ within NorCPM, a fully coupled Earth system model~~– in both online and offline scenarios~~ , to improve seasonal sea ice predictions in the Arctic under both online and offline scenarios. In the ~~context of~~ online error correction approach, ML is utilized to rectify errors in instantaneous model states in the middle of the month ~~, providing instantaneous corrections~~ that serve as initial conditions for the subsequent model integration. The offline error correction approach involves the post-processing of monthly sea ice predictions.

The approaches proposed in this study integrate ML with a dynamical modeling framework, with the primary objective of reducing the intrinsic prediction errors of the dynamical model itself. Unlike purely data-driven models (e.g., Andersson et al., 2021; Ren et , which are typically designed for statistical prediction of specific sea ice properties, ML here aims to improve the overall performance of the dynamical prediction system that ensures physical consistency among a large number of predicted variables.

Our results demonstrate that both ~~the~~ online and offline ML-based error correction models can well predict the spatial distribution of errors, albeit with slight deficiencies in capturing amplitude. By applying the two approaches to seasonal Arctic sea ice predictions initialized from January, April, July, and October, we ~~found~~ find that both approaches can reduce SIE and

IIEE prediction errors compared to the raw predictions without error correction. ~~Moreover, the~~ The improvements vary with the lead month, ~~e.g., significant improvements~~ with particularly notable enhancements observed in predictions from August to October.

~~Comparing~~ By comparing the two error correction approaches, we ~~found~~ find that the offline approach ~~overall~~ yields smaller errors than the online approach. ~~This may be because the online~~ The online error correction approach corrects instantaneous model errors ~~(observed~~ only on the 15th day of the month~~). However, during the model integration, the impact of error correction gradually dissipates due to~~ , 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, ~~when computing monthly averaged outputs, the impact of the error correction is damped. It could be beneficial to apply the online error correction model more frequently in time (Gregory et al., 2024). However, the training would not be as straightforward since the analysis increment of NorCPM is only available once per month. Alternatively, training the hybrid model as a whole was beneficial (Farchi et al., 2021)but it requires the external constraint~~ it is computationally more efficient and easier to integrate into operational sea ice prediction systems than the online approach.

It is important to note that the proposed approaches still have room for improvement. In this study, we only use ocean and sea ice variables as input features. Including atmospheric variables would help to address errors due to both dynamic and thermodynamic processes and further improve the performance. Increasing the frequency of online correction could help enhance its effectiveness (Gregory et al., 2024), but this is challenging in practice since analysis increments in NorCPM are currently available only every month. An alternative strategy is to train hybrid models that combine ML with dynamical models, which has been shown to be effective in other systems (Farchi et al., 2021). However, this approach relies on external constraints to compute the gradient of the dynamical model, which ~~was~~ are not available in ~~our case. In contrast, the offline approach is designed to directly correct the monthly mean output without the need for model integration.~~ NorCPM.

~~When examining the improvements in regional SIE or SIC, the most significant improvements are observed near the ice edge where sea ice dynamics are active. Overall, the error correction schemes demonstrate their effectiveness in these regions, particularly during the periods when the sea ice dynamics are most pronounced and NorCPM exhibits large errorstypically from September to November. During these months, the sea ice margins are subject to rapid changes, and~~ 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 ~~error correction approaches can capture and adjust for these variations accurately, leading to better model performance in these critical regions.~~ 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).

~~Our error correction schemes operate on a grid-point basis. As mentioned before, our ML model does not utilize spatial patterns, which explains some of the limitations of our approaches, particularly when the NorCPM's raw hindcasts are already accurate. However, this simplicity offers considerable flexibility in applying the error correction models and reduces the risk of overfitting to specific spatial patterns.~~

Code and data availability. The code and data to plot figure is available at https://doi.org/10.5281/zenodo.14533027.

Author contributions. Conceptualization: ZH, YW, JB. Analysis and Visualization: ZH. Interpretation of results: ZH, YW, JB. Writing (original draft): ZH, YW. Writing (reviewing and editing original draft): ZH, YW, JB, XW, ZS.

Competing interests. The author declares that no competing interests.

Acknowledgements. This study was funded by National Natural Science Foundation of China (42225602), the National Key $R\&D$ Program of China (2022YFE0106400), ~~the China Scholarship Council (202206710071),~~ Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX23_0657), the China Scholarship Council (202206710071), the Special Founds for Creative Research (2022C61540), the Opening Project of the Key Laboratory of Marine Environmental Information Technology (521037412). JB was funded by the project TARDIS funded by the Research Council of Norway ~~grant number 325241.~~ (under grant No. 325241). YW was funded by the Norges Forskningsråd (under grant No. 328886) and the Trond Mohn stiftelse (under grand No. BFS2018TMT01). This work also received grants for computer time from the Norwegian Program for supercomputer (NN9039K) and storage grants (NS9039K). ZS was funded by the National Natural Science Foundation of China (42176003).

References

Andersson, T. R., Hosking, J. S., Pérez-Ortiz, M., Paige, B., Elliott, A., Russell, C., Law, S., Jones, D. C., Wilkinson, J., Phillips, T., et al.: Seasonal Arctic sea ice forecasting with probabilistic deep learning, Nature communications, 12, 5124, 2021.

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.

Bethke, I., Wang, Y., Counillon, F., Keenlyside, N., Kimmritz, M., Fransner, F., Samuelsen, A., Langehaug, H., Svendsen, L., Chiu, P.-G., et al.: NorCPM1 and its contribution to CMIP6 DCPP, Geoscientific Model Development Discussions, 2021, 1–84, 2021.

Blanchard-Wrigglesworth, E., Barthélemy, A., Chevallier, M., Cullather, R., Fučkar, N., Massonnet, F., Posey, P., Wang, W., Zhang, J., Ardilouze, C., et al.: Multi-model seasonal forecast of Arctic sea-ice: forecast uncertainty at pan-Arctic and regional scales, Climate Dynamics, 49, 1399–1410, 2017.

Bleck, R., Dean, S., O'Keefe, M., and Sawdey, A.: A comparison of data-parallel and message-passing versions of the Miami Isopycnic Coordinate Ocean Model (MICOM), Parallel computing, 21, 1695–1720, 1995.

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.

Carrassi, A., Weber, R., Guemas, V., Doblas-Reyes, F., Asif, M., and Volpi, D.: Full-field and anomaly initialization using a low-order climate model: a comparison and proposals for advanced formulations, Nonlinear Processes in Geophysics, 21, 521–537, 2014.

Carrassi, A., Bocquet, M., Bertino, L., and Evensen, G.: Data assimilation in the geosciences: An overview of methods, issues, and perspectives, Wiley Interdisciplinary Reviews: Climate Change, 9, e535, 2018.

Counillon, F., Bethke, I., Keenlyside, N., Bentsen, M., Bertino, L., and Zheng, F.: Seasonal-to-decadal predictions with the ensemble Kalman filter and the Norwegian Earth System Model: a twin experiment, Tellus A: Dynamic Meteorology and Oceanography, 66, 21 074, 2014.

Counillon, F., Keenlyside, N., Bethke, I., Wang, Y., Billeau, S., Shen, M. L., and Bentsen, M.: Flow-dependent assimilation of sea surface temperature in isopycnal coordinates with the Norwegian Climate Prediction Model, Tellus A: Dynamic Meteorology and Oceanography, 68, 32 437, 2016.

Craig, A. P., Vertenstein, M., and Jacob, R.: A new flexible coupler for earth system modeling developed for CCSM4 and CESM1, The International Journal of High Performance Computing Applications, 26, 31–42, 2012.

De Cruz, L., Demaeyer, J., and Vannitsem, S.: The modular arbitrary-order ocean-atmosphere model: MAOOAM v1. 0, Geoscientific Model Development, 9, 2793–2808, 2016.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., Bauer, d. P., et al.: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, Quarterly Journal of the royal meteorological society, 137, 553–597, 2011.

Evensen, G.: The ensemble Kalman filter: Theoretical formulation and practical implementation, Ocean dynamics, 53, 343–367, 2003.

Farchi, A., Laloyaux, P., Bonavita, M., and Bocquet, M.: Using machine learning to correct model error in data assimilation and forecast applications, Quarterly Journal of the Royal Meteorological Society, 147, 3067–3084, 2021.

Gent, P. R., Danabasoglu, G., Donner, L. J., Holland, M. M., Hunke, E. C., Jayne, S. R., Lawrence, D. M., Neale, R. B., Rasch, P. J., Vertenstein, M., et al.: The community climate system model version 4, Journal of climate, 24, 4973–4991, 2011.

Goessling, H. F., Tietsche, S., Day, J. J., Hawkins, E., and Jung, T.: Predictability of the Arctic sea ice edge, Geophysical Research Letters, 43, 1642–1650, https://doi.org/https://doi.org/10.1002/2015GL067232, 2016.

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.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, 2020.

Heuzé, C. and Jahn, A.: The first ice-free day in the Arctic Ocean could occur before 2030, Nature Communications, 15, 1–10, 2024.

Holland, M. M., Bailey, D. A., Briegleb, B. P., Light, B., and Hunke, E.: Improved sea ice shortwave radiation physics in CCSM4: The impact of melt ponds and aerosols on Arctic sea ice, Journal of Climate, 25, 1413–1430, 2012.

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.

Jung, T., Gordon, N. D., Bauer, P., Bromwich, D. H., Chevallier, M., Day, J. J., Dawson, J., Doblas-Reyes, F., Fairall, C., Goessling, H. F., et al.: Advancing polar prediction capabilities on daily to seasonal time scales, Bulletin of the American Meteorological Society, 97, 1631–1647, 2016.

Kim, Y. J., Kim, H.-c., Han, D., Stroeve, J., and Im, J.: Long-term prediction of Arctic sea ice concentrations using deep learning: Effects of surface temperature, radiation, and wind conditions, Remote Sensing of Environment, 318, 114 568, 2025.

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.

Kirkevåg, A., Grini, A., Olivié, D., Seland, Ø., Alterskjær, K., Hummel, M., Karset, I. H., Lewinschal, A., Liu, X., Makkonen, R., et al.: A production-tagged aerosol module for Earth system models, OsloAero5. 3–extensions and updates for CAM5. 3-Oslo, Geoscientific Model Development, 11, 3945–3982, 2018.

Laloyaux, P., de Boisseson, E., Balmaseda, M., Bidlot, J.-R., Broennimann, S., Buizza, R., Dalhgren, P., Dee, D., Haimberger, L., Hersbach, H., et al.: CERA-20C: A coupled reanalysis of the twentieth century, Journal of Advances in Modeling Earth Systems, 10, 1172–1195, 2018.

Lawrence, D. M., Oleson, K. W., Flanner, M. G., Thornton, P. E., Swenson, S. C., Lawrence, P. J., Zeng, X., Yang, Z.-L., Levis, S., Sakaguchi, K., et al.: Parameterization improvements and functional and structural advances in version 4 of the Community Land Model, Journal of Advances in Modeling Earth Systems, 3, 2011.

Onarheim, I. H., Eldevik, T., Smedsrud, L. H., and Stroeve, J. C.: Seasonal and regional manifestation of Arctic sea ice loss, Journal of Climate, 31, 4917–4932, 2018.

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.

Penny, S. G. and Hamill, T. M.: Coupled data assimilation for integrated earth system analysis and prediction, Bulletin of the American Meteorological Society, 98, ES169–ES172, 2017.

Ren, Y., Li, X., and Wang, Y.: SICNet season V1. 0: a transformer-based deep learning model for seasonal Arctic sea ice prediction by integrating sea ice thickness data, Geoscientific Model Development Discussions, 2024, 1–20, 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.

Saha, S., Nadiga, S., Thiaw, C., Wang, J., Wang, W., Zhang, Q., Van den Dool, H., Pan, H.-L., Moorthi, S., Behringer, D., et al.: The NCEP climate forecast system, Journal of Climate, 19, 3483–3517, 2006.

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.

Serreze, M. C., Holland, M. M., and Stroeve, J.: Perspectives on the Arctic's shrinking sea-ice cover, science, 315, 1533–1536, 2007.

Stroeve, J. C., Markus, T., Boisvert, L., Miller, J., and Barrett, A.: Changes in Arctic melt season and implications for sea ice loss, Geophysical Research Letters, 41, 1216–1225, 2014.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An overview of CMIP5 and the experiment design, Bulletin of the American meteorological Society, 93, 485–498, 2012.

Thornton, E.: Technical Description of version 4.0 of the Community Land Model (CLM), NCAR, Climate and Global, 2010.

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.

Wagner, P. M., Hughes, N., Bourbonnais, P., Stroeve, J., Rabenstein, L., Bhatt, U., Little, J., Wiggins, H., and Fleming, A.: Sea-ice information and forecast needs for industry maritime stakeholders, Polar Geography, 43, 160–187, 2020.

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, X., Liu, Y., Key, J. R., and Dworak, R.: A new perspective on four decades of changes in Arctic sea ice from satellite observations, Remote Sensing, 14, 1846, 2022.

Wang, Y., Counillon, F., and Bertino, L.: Alleviating the bias induced by the linear analysis update with an isopycnal ocean model, Quarterly Journal of the Royal Meteorological Society, 142, 1064–1074, 2016.

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.

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.

Watt-Meyer, O., Brenowitz, N. D., Clark, S. K., Henn, B., Kwa, A., McGibbon, J., Perkins, W. A., and Bretherton, C. S.: Correcting weather and climate models by machine learning nudged historical simulations, Geophysical Research Letters, 48, e2021GL092 555, 2021.

Weber, R. J., Carrassi, A., and Doblas-Reyes, F. J.: Linking the anomaly initialization approach to the mapping paradigm: a proof-of-concept study, Monthly Weather Review, 143, 4695–4713, 2015.

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.

Zuo, H., Balmaseda, M. A., Tietsche, S., Mogensen, K., and Mayer, M.: The ECMWF operational ensemble reanalysis–analysis system for ocean and sea ice: a description of the system and assessment, Ocean science, 15, 779–808, 2019.