Deep learning of systematic sea ice model errors from data assimilation increments

William Gregory[1], Mitchell Bushuk[2], Alistair Adcroft[1], Yongfei Zhang[1], Laure Zanna[3]

[1]Atmospheric and Oceanic Sciences Program, Princeton University, NJ, USA
[2]Geophysical Fluid Dynamics Laboratory, NOAA, Princeton, NJ, USA
[3]Courant Institute of Mathematical Sciences, New York University, New York, NY, USA

arXiv:2304.03832v1 [physics.ao-ph] 7 Apr 2023

Key Points:

- We show that sea ice data assimilation increments closely reflect the systematic bias patterns of a global ice-ocean model
- Convolutional neural networks can make skillful predictions of sea ice data assimilation increments, using only model state variables
- The skillful predictions suggest the network could be used as a parameterization to reduce sea ice biases in free-running model simulations

Corresponding author: Will Gregory, wg4031@princeton.edu

Abstract

Data assimilation is often viewed as a framework for correcting short-term error growth in dynamical climate model forecasts. When viewed on the time scales of climate however, these short-term corrections, or analysis increments, can closely mirror the systematic bias patterns of the dynamical model. In this study, we use convolutional neural networks (CNNs) to learn a mapping from model state variables to analysis increments, in order to showcase the feasibility of a data-driven model parameterization which can predict state-dependent model errors. We undertake this problem using an ice-ocean data assimilation system within the Seamless system for Prediction and EArth system Research (SPEAR) model, developed at the Geophysical Fluid Dynamics Laboratory, which assimilates satellite observations of sea ice concentration every 5 days between 1982–2017. The CNN then takes inputs of data assimilation forecast states and tendencies, and makes predictions of the corresponding sea ice concentration increments. Specifically, the inputs are states and tendencies of sea ice concentration, sea-surface temperature, ice velocities, ice thickness, net shortwave radiation, ice-surface skin temperature, sea-surface salinity, as well as a land-sea mask. We find the CNN is able to make skillful predictions of the increments in both the Arctic and Antarctic and across all seasons, with skill that consistently exceeds that of a climatological increment prediction. This suggests that the CNN could be used to reduce sea ice biases in free-running SPEAR simulations, either as a sea ice parameterization or an online bias correction tool for numerical sea ice forecasts.

Plain Language Summary

To make predictions of the Earth's climate system we use expensive computer simulations, called climate models. These models are not perfect however, as we often need to approximate certain physical laws in order to save on compute time. On the other hand we have observational climate data, however these data have limited space and time coverage and also contain errors because of noise and assumptions about how our measurements relate to the quantity we are interested in. Therefore we often use a process called data assimilation to combine our climate model predictions together with observations, to produce our 'best guess' of the climate system. The difference between our best-guess-model and our original climate model prediction then gives us clues as to how wrong our original climate model is. In this work we use some fancy statistics, called machine learning, where we show a computer algorithm lots of examples of sea ice, atmosphere and ocean climate model predictions, and see if it can learn its own inherent sea ice errors. We find that it can do this well, which means that we can hopefully incorporate the machine learning algorithm into the original climate model to improve its future climate predictions.

1 Introduction

The influence of structural errors within climate models due to missing physics, imperfect parameterizations of subgrid-scale processes, as well as errors in the underlying numerics, leads to systematic biases across the atmosphere, land, sea ice, and ocean. Subsequently, our ability to diagnose and correct these biases ultimately governs the accuracy of numerical weather and climate predictions on different time scales (Stevens & Bony, 2013). In the context of sea ice for example, much effort has been afforded to the improvement of model physics and subgrid parameterizations through the development of e.g., ice thickness distribution (Thorndike et al., 1975; Bitz et al., 2001) and floe-size distribution theory (Rothrock & Thorndike, 1984; Horvat & Tziperman, 2015), surface melt-pond (Flocco et al., 2012), ice drift (Tsamados et al., 2013) and lateral melt parameterizations (M. Smith et al., 2022), as well as sea ice rheology (Hibler, 1979; Dansereau et al., 2016; Ólason et al., 2022). Such studies have shown how the improved represen-

tation of sea ice physics produces model simulations which more closely reflect observations in terms of either their mean sea ice volume, drift, or ice thickness distribution. Despite this, however, biases will often persist due to the fact that physical processes must be approximated in order to meet computational restraints, and that parameterizations are often based on sparse observations which were collected under a climate regime which may not generalize to future conditions (Notz, 2012). Sea ice is also strongly coupled to both the atmosphere and ocean via mechanical and thermodynamic forcing, thus sea ice biases can also manifest from biases in these components.

Many previous studies have leveraged data assimilation (DA) as a way to either assess model error or better understand model physics within numerical weather prediction (NWP) systems (Leith, 1978; Klinker & Sardeshmukh, 1992; Dee, 2005; Rodwell & Palmer, 2007; Palmer & Weisheimer, 2011; Carrassi & Vannitsem, 2011; Mitchell & Carrassi, 2015; Crawford et al., 2020; Laloyaux et al., 2020). Generally, DA can be considered a Bayesian framework for combining a model forecast with observations in order to produce an optimal estimate of a given set of climate state variables, often called the *analysis state*. The difference between this analysis state and the model forecast prior to assimilation is then the *analysis increment*, which represents our 'best guess' as to the appropriate correction to the model forecast when taking into account both model and observational uncertainty. One caveat to this is that many DA systems do not formally account for systematic model biases, and so these systems often produce non-zero values in the time-mean of their analysis increments; indicating consistent discrepancies between the model and observations. Attributing such errors to their correct source is also non-trivial (Dee, 2004, 2005), as model biases can manifest non-locally in space and time (Palmer & Weisheimer, 2011; C. Wang et al., 2014) and involve non-linear interactions across different model components (Large & Danabasoglu, 2006; Kim et al., 2022). Observations themselves may also contain systematic errors, such as the design of weather filters in satellite-derived sea ice area retrievals (Kern et al., 2019) and uncertainties related to summer ice surface properties (Kern et al., 2020). While some studies have shown relative success in separating systematic errors between observations and models (Auligné et al., 2007; Dee & Uppala, 2009), many assimilation systems simply assume that the observational errors are uncorrelated and Gaussian, and subsequently any systematic patterns within the analysis increments can largely be considered a manifestation of the various model biases. Under this assumption, the increments can be seen as a reflection of model error growth associated with missing or imbalanced physical processes occurring over short time scales, often called *fast physics* errors, however such errors ultimately have an impact on the model's bias patterns over climate time scales as well (J. M. Murphy et al., 2004; Rodwell & Palmer, 2007).

The analysis increments therefore provide useful information on model deficiencies, which could inform new parameterizations to reduce systematic model biases. Indeed, variational schemes such as weak-constraint 4D-Var (Wergen, 1992; Zupanski, 1993; Trémolet, 2007; Laloyaux et al., 2020) already aim to account for systematic model error during DA, and while this is invaluable in NWP, the underlying model physics remains unchanged, meaning that a free-running model simulation invariably remains biased. An alternative approach which has been explored in the ocean modeling community (Chepurin et al., 2005; Balmaseda et al., 2007; Lu et al., 2020) is to use DA to first derive the climatological components of the systematic model biases, and then incorporate these components back into the model as an adjustment to the model state tendencies. Lu et al. (2020) for example designed an ocean DA system which assimilates temperature and salinity profile data into version 6 of the Modular Ocean Model (MOM6), and from this they derived analysis increments of temperature and salinity at each model grid cell location and vertical level. They subsequently computed the daily climatology of the increments, which represent the systematic component of model error for each field on any given day of the year, and incorporated these as a three-dimensional adjustment to the model temperature and salinity tendencies for subsequent MOM6 ocean simulations. This 'ocean ten-

dency adjustment' was found to reduce ocean model bias and improve the skill of coupled model seasonal predictions of the El Niño Southern Oscillation.

More recently, machine learning (ML) has been put forward as a data-driven framework for targeting model biases. ML, in particular deep learning (DL), algorithms have become increasingly popular in climate research for a variety of applications ranging from NWP (Pathak et al., 2022; Bi et al., 2022) to satellite altimetry data processing (Dawson et al., 2022; Landy et al., 2022). In the context of dynamical climate models, DL algorithms have proven effective tools for deriving model parameterizations directly from numerical simulations. For example, many past studies have focused on learning subgrid parameterizations from high resolution experiments and/or observations of the ocean (Bolton & Zanna, 2019; Zanna & Bolton, 2020; Zhu et al., 2022), atmosphere (Brenowitz & Bretherton, 2018; Gentine et al., 2018; Rasp et al., 2018; O'Gorman & Dwyer, 2018; Yuval & O'Gorman, 2020; P. Wang et al., 2022), and sea ice (Finn et al., 2023). In the context of DA-based approaches, some recent studies have relied on iterative sequences of DA and ML to infer unresolved scale parameterizations from sparse and noisy observations (Brajard et al., 2021), or to learn state-dependent model error from analysis increments (Farchi et al., 2021) and nudging tendencies (Watt-Meyer et al., 2021; Bretherton et al., 2022), while others have combined DA with equation discovery to extract interpretable structural model errors (Mojgani et al., 2022). Many of these studies have relied on idealized models to showcase the feasibility of various DA-ML methodologies, however recently Bonavita and Laloyaux (2020) used ML to learn state-dependent model errors from atmospheric analysis increments produced from a 4D-Var simulation within the the Integrated Forecasting System (IFS) model at the European Centre for Medium-range Weather Forecasts (ECMWF), and similarly, Laloyaux et al. (2022) attempted to learn atmospheric temperature errors within the same IFS model using the model bias directly, as a way to a-priori define the bias model within subsequent 4D-Var simulations. This latter approach however was unable to outperform the current operational weak-constraint 4D-Var system at ECMWF.

In this study, we present a DA-based ML approach to learn the systematic biases of a large-scale sea ice model used for climate simulations. We learn state-dependent sea ice errors within the Seamless system for Prediction and EArth system Research (SPEAR) model (Delworth et al., 2020), developed at the Geophysical Fluid Dynamics Laboratory (GFDL), by constructing convolutional neural networks (CNNs) which learn a functional mapping from model state variables to sea ice DA increments. Somewhat different to previous studies which have been centered around DA and ML in idealized model contexts (Brajard et al., 2021; Farchi et al., 2021; Mojgani et al., 2022), our application here is, to our knowledge, the first example of using ML to learn systematic model error from DA increments in a global ice-ocean model (though similar approaches have previously been explored within large-scale atmospheric models (Bonavita & Laloyaux, 2020; Chen et al., 2022)). We also choose to learn sea ice errors from DA increments as opposed to learning the model bias directly (e.g., Laloyaux et al. (2022)), as the increments have inherently accounted for model and observational uncertainty, and they also provide a full spatio-temporal record of errors for model state variables which are not direct observables, such as subgrid ice thickness distribution category concentrations. It is also worth noting that while we present this article in the context of using ML to make offline predictions of sea ice DA increments, we are ultimately working towards an ML model which can be implemented as an online sea ice parameterization within SPEAR. Similar to previous works (Grundner et al., 2022; P. Wang et al., 2022), this article is therefore an initial evaluation into the feasibility of this task, based on offline performance.

This paper is structured as follows: Section 2 provides a brief overview of the SPEAR ice-ocean model configuration, as well as the sea ice DA setup. Section 3 then highlights how the climatological sea ice concentration (SIC) bias of a SPEAR ice-ocean model experiment maps closely onto the SPEAR SIC DA increments, motivating the idea of learn-

ing systematic model error from analysis increments. Section 4 describes the ML problem setup and documents the CNN architectures and hyperparameter settings. Section 5 then showcases the predictive performance of the CNN, and provides an assessment of the CNN sensitivity and generalization ability. Section 6 presents a discussion on the results and outlines considerations for future work relating to sea ice parameterizations and climate prediction. A final summary is then given in section 7, as well as an outlook on the broader implications of this work within the climate modeling community.

2 Model configuration

2.1 SPEAR ice-ocean model

SPEAR is a fully coupled ice-ocean-atmosphere-land model, with nominal 1° horizontal resolution in the ice and ocean components (Delworth et al., 2020). The SPEAR ocean component is based on MOM6, with 75 vertical layers, and the sea ice component on version 2 of the Sea Ice Simulator (SIS2; see Adcroft et al. (2019) for details on both MOM6 and SIS2). In this work, we consider an ice-ocean model configuration of SPEAR forced by atmospheric conditions and river runoff from the Japanese 55-year Reanalysis for driving ocean-sea-ice models (JRA55-do; Tsujino et al. (2018)).

The SIS2 ice dynamics are solved using a elastic-viscous-plastic rheology on a tripolar Arakawa C-grid (Bouillon et al., 2009), with advection performed with a modified upwind scheme (Adcroft et al., 2019). The energy-conserving thermodynamics of the ice follows that of Bitz and Lipscomb (1999), and uses a vertical structure consisting of four ice layers and a single snow layer. Following Bitz et al. (2001), five ice thickness distribution categories are implemented in a Lagrangian scheme, with thickness boundaries of 0.1, 0.3, 0.7, 1.1 metres. The coupling between ice and ocean occurs at a frequency of 60 minutes, with a temperature coupling coefficient of 240 $\mathrm{Wm^{-2}K^{-1}}$, while faster coupling with the atmosphere occurs through a surface skin temperature every 20 minutes. The model does not contain melt-pond, subgrid ridging, lateral melt, or land-fast ice parameterizations.

2.2 Sea ice data assimilation and model experiments

An experimental ice-ocean DA system within SPEAR was recently developed by Y. Zhang et al. (2021), whereby satellite-derived SIC from the National Snow and Ice Data Center (NSIDC; Cavalieri et al. (1996)) NASA Team algorithm is assimilated into SIS2 via the Ensemble Adjustment Kalman Filter (EAKF; Anderson (2001)), and MOM6 sea-surface temperatures are nudged towards observations from version 2 of the 1° gridded Optimum Interpolation Sea-Surface Temperature (OISSTv2) data set (Reynolds et al., 2007; Banzon et al., 2016). In this section we give a brief overview of this DA setup, although the reader is referred to Y. Zhang et al. (2021) for further details.

A single SPEAR ensemble member is initialised in 1958 with World Ocean Atlas ocean conditions, and a prescribed atmosphere from JRA55-do reanalysis. This single member is then integrated forward to 1979 in order to 'spin up' the ocean and sea ice, which then provides the initial ice and ocean conditions for a set of 30 ensemble members, each with individual perturbed sea ice physics. These perturbations correspond to independent random draws from a uniform distribution for sea ice model parameters including the ice strength parameter (Hibler, 1979), as well as the ice, snow, and pond albedo parameters (Briegleb & Light, 2007). The distribution for ice strength spans 20,000–50,000 $\mathrm{Nm^{-1}}$, while the distribution for albedo parameters spans -1.6–1.6 standard deviations (Y. Zhang et al., 2021). The 30 perturbed physics ensemble members are then integrated forward from 1979 to 1982 in order to spin up the sea ice and generate sufficient spread across the ensemble. After which, the first sea ice DA update is made on January 6$^{\mathrm{th}}$ 1982 and continues every 5 days until December 27$^{\mathrm{th}}$ 2017, providing a total of 2618 as-

similation cycles. This corresponds to 73 cycles per year except for 1982, 1987 and 1988, which contain 71, 68 and 70 cycles, respectively. There are 71 cycles in 1982 because the first cycle begins after the initial update on January 6$^{\text{th}}$, and 68 and 70 in 1987 and 1988 due to missing satellite observations between December 3$^{\text{rd}}$ 1987 and January 13$^{\text{th}}$ 1988 (Cavalieri et al., 1996). Note that, for convenience, the model is run with a 'no leap' calendar which excludes leap-year days.

During each assimilation cycle, a model forecast is run until 00:00 hours UTC on the assimilation day (e.g., Jan 6$^{\text{th}}$), at which point the ice concentration from each of the model's five individual ice thickness distribution categories (hereafter SICN; note that $\text{SIC} = \sum_{k=1}^{5} \text{SICN}_k$) are passed to the EAKF, along with the satellite SIC observations, and subsequently the SICN forecasts are updated by the filter to produce their analysis states. Given that the aggregate SIC analysis state corresponds to the sum of the SICN analysis states, it is necessary to post-process SICN after each DA cycle in order avoid non-physical values in SIC, which is bounded between 0 and 1. This is achieved by appropriately scaling each of the SICN states when SIC is greater than 1, and setting SICN to 0 when SIC is negative. After post-processing, the analysis increments are then computed for each of the five category concentrations (ΔSICN), and for each of the 30 ensemble members. State variables for each ensemble member are saved as daily mean fields during model integration, giving 365 days $\times$ 36 years = 13140 daily forecasts for each variable. For the remainder of this article we consider only the ensemble mean fields for both the model state variables and the analysis increments.

In order to understand the inherent SIC bias patterns within SPEAR, the next section includes a comparison of the SIC DA increments (ΔSIC) to an additional model experiment without SIC DA, referred to here as FREE. This experiment corresponds to the same JRA-forced ice-ocean model configuration with sea-surface temperature nudging, as well as the same perturbed sea ice physics, and initial conditions from the spinup run as the SIC DA experiment. Therefore the FREE experiment configuration is identical to the SIC DA run, except for the assimilation of SIC observations.

3 Analysis increments and model bias in SPEAR

Learning systematic model error from DA increments, with the goal of an eventual sea ice parameterization which reduces climate model bias, relies on the assumption that the fast physics errors captured within the DA increments reflect the long-term systematic biases of the free-running model (Rodwell & Palmer, 2007). In this section, we examine whether this necessary condition is satisfied, making comparisons of ΔSIC to the climatological bias of the FREE experiment. The model bias is computed relative to NSIDC NASA Team satellite SIC observations.

Figure 1 shows seasonal climatologies of the SPEAR FREE SIC model bias and ΔSIC between 1982–2017, for both the Arctic and Antarctic. Here we notice that the free-running model is, on average, positively biased in both hemispheres, with larger magnitude biases in the Antarctic. Crucially, we find largely consistent patterns between the model bias and ΔSIC. In the Arctic for example, the large positive biases in the Greenland, Iceland, Norwegian (GIN) and Barents seas (east Atlantic) are mirrored by overall negative increments, hence the DA is acting to remove sea ice in this region. The winter Arctic SIC biases appear to be related to systematic biases in the sea ice edge position, which is apparent when noticing that the increments in the fully covered ice pack (north of the 75% observed SIC contour) are relatively small compared to the marginal ice zones in DJF and MAM. The presence of larger increments in the central ice pack in JJA and SON are then likely a reflection of local SIC errors in the ice-covered zone in addition to ice edge position errors. The only notable discrepancy between model bias and ΔSIC in the Arctic appears to be in the Kara and Laptev shelf seas in JJA, where both the model bias and increments are positive. This suggests that the assimilation fore-

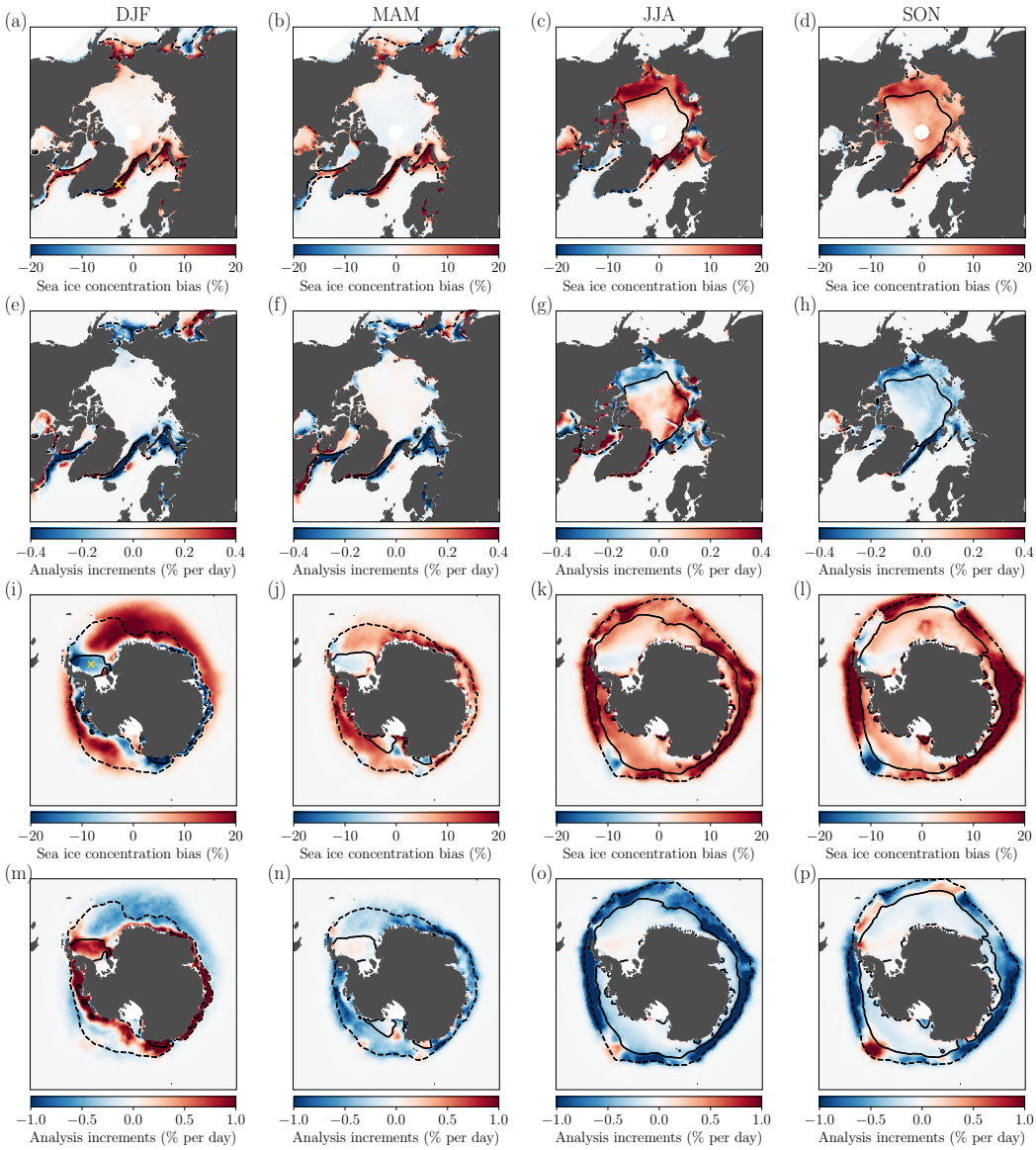

Figure 1. Seasonal climatologies of SPEAR free-running model bias (model minus observations) and sea ice concentration analysis increments, for both the Arctic (a)–(h) and Antarctic (i)–(p). Columns from left to right show DJF, MAM, JJA, SON climatologies, computed over the period 1982–2017. Dashed and solid contours denote the observed climatology marginal ice zone boundaries over the same period (15 and 75% SIC contours, respectively). Yellow markers in (a) and (i) are example grid-point locations used for analysis in Figure 2.

casts are negatively biased in this region, which may be related to a residual overshooting problem in the DA experiment, as highlighted in the original SPEAR sea ice DA study by Y. Zhang et al. (2021).

Turning to the Antarctic, despite largely positive biases across all seasons, negative biases dominate many of the coastal regions in the austral summer (DJF), including the Weddell Sea, whereby many of these biases become lower in magnitude or even positive by austral winter (JJA). Interestingly, the isolated negative bias towards the north-

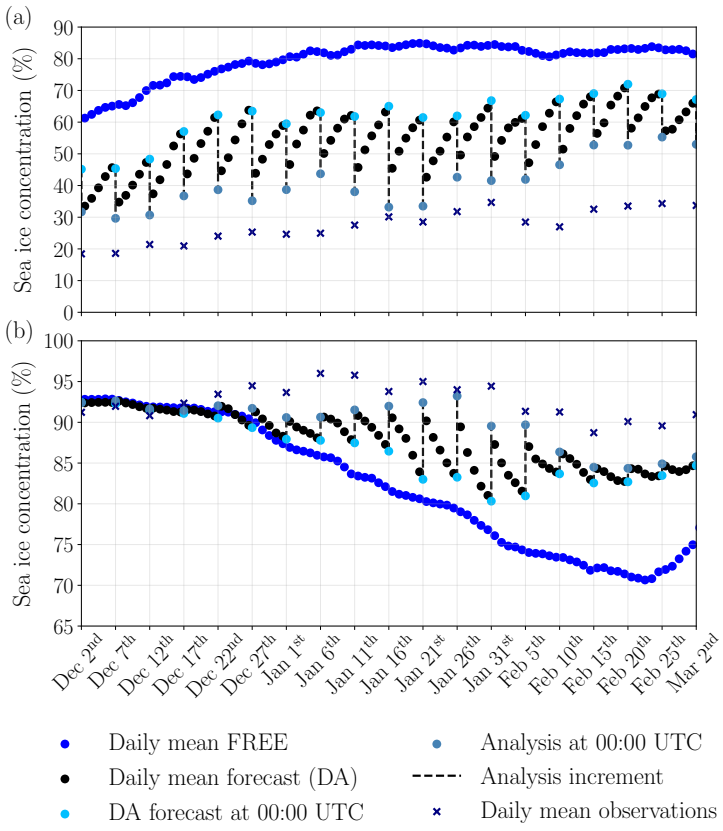

Figure 2. SPEAR sea ice concentration data assimilation example, shown for one grid cell as daily climatologies (1982–2017). Examples are presented for the Arctic (a) and Antarctic (b) through the period December–February. The grid cells for both the Arctic and Antarctic examples correspond to locations in the GIN Sea and Weddell Sea, respectively (see the yellow markers in Figures 1a and 1i).

eastern edge of the Ross Sea is a persistent feature from MAM through to SON, reaching its largest magnitude in SON. This may be related to strong deep ocean convection in this region (Adcroft et al., 2019), which manifests as positively biased sea-surface temperatures which are co-located with the negatively biased SIC zone (see Figure S1 in Supporting Information S1). Overall, the strong spatial and seasonal agreement between the free-running model bias and ΔSIC supports this study's plan to use DA increments to learn a parameterization of sea ice model error.

Visualising the time evolution of the sea ice DA forecasts (Figure 2) shows the relationship between systematic biases and analysis increments more clearly. In the GIN Sea (Figure 2a), we can see that the model forecasts in each DA cycle (black dots) are drifting towards the positively-biased free-running model state (dark blue dots) over the 5-day forecast period, and as such the analysis increments (dashed black lines) are systematically negative to account for this. Similarly, in the Weddell Sea (Figure 2b) the forecasts are drifting towards the negatively-biased free-running model state, resulting in systematically positive increments. The forecast drift that is observed in either case can be quantified by the assimilation forecast tendencies, which for a given assimilation cycle i, corresponds to the time-derivative of the forecast c at time t, or more simply $\dot{c}_i(t) = c_i(t) - c_i(t-1)$. The *total* forecast tendency for a given assimilation cycle is then the

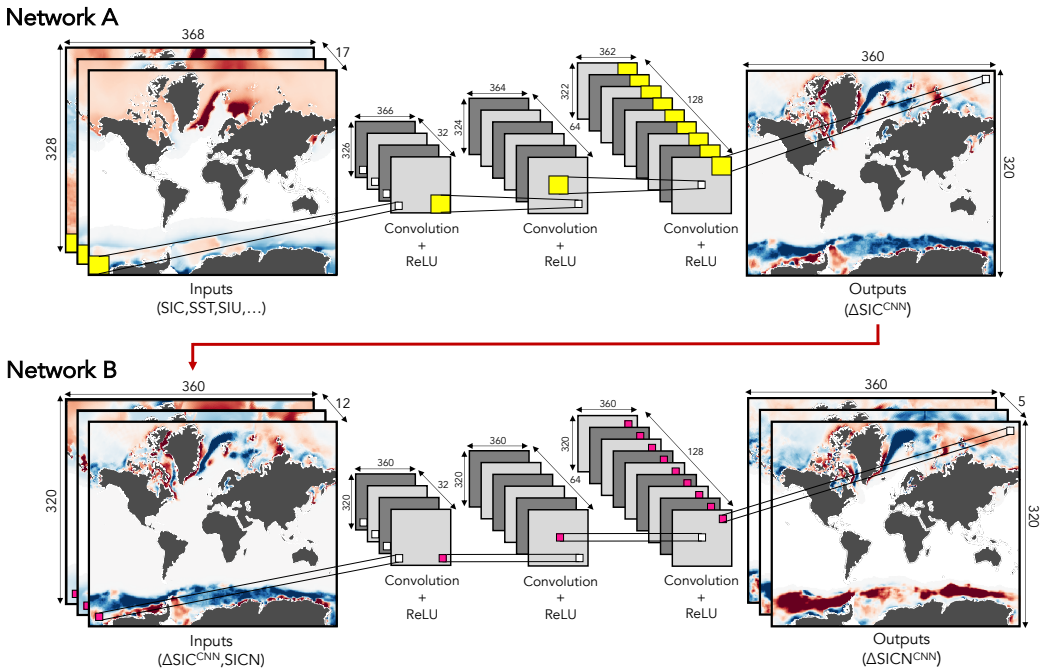

Figure 3. Schematic of the CNN architectures used to learn functional mappings from state vectors to analysis increments. The yellow and purple squares represent 3×3 and 1×1 kernels over which the convolution operations are performed in each layer, respectively, where there is one kernel for every feature map in each layer. The white pixel is then the sum of convolution outputs from all features in the previous layer, which has subsequently been passed through a ReLU activation function. The activation function after the last convolution operation to the output layer is the identity function.

sum of the individual daily tendencies: $\dot{c}_i(1)+\dot{c}_i(2)+...+\dot{c}_i(5)$. Klinker and Sardeshmukh (1992) showed that the mean total tendencies across a large number of assimilation cycles, referred to as the *systematic forecast tendency*, is approximately equal to the negative of the analysis increments, which is also the case in our SPEAR DA experiments (see Figure S2). Building on this, Rodwell and Palmer (2007) then later described how the forecast tendencies can be broken down into tendencies associated with the model's representation of various resolved and parameterized physical processes, and subsequently used them to make assessments of model physics errors after a model change had been made. In our study here, we utilize this inherent link between forecast tendencies and analysis increments to construct CNNs which use inputs of both state variables from the DA forecasts, as well as their associated forecast tendencies, in order to predict ΔSICN.

4 Convolutional neural networks

CNNs are a specific class of DL algorithms which are well-suited to problems where inputs contain local correlation structure in space and/or time (K. Murphy, 2022). For this reason they have historically been successful in the domains of image recognition and segmentation (Simonyan & Zisserman, 2014; Zeiler & Fergus, 2014; Dong et al., 2015; Ronneberger et al., 2015; Krizhevsky et al., 2017), where the aim is to e.g., classify objects or isolate features within medical images. In Earth system modeling CNNs have subsequently been utilized for their ability to exploit the two-dimensional structure associated with turbulent fluids, and hence learn subgrid parameterizations of ocean mesoscale

Table 1. Details of the convolutional neural networks (CNNs) inputs, outputs, architecture, and hyperparameters used during training.

	Network A	**Network B**
Inputs (* states & tendencies)	SIC*, SST*, SIU*, SIV*, SIT*, SW*, TS*, SSS*, Land-sea mask	$\Delta\text{SIC}^{\text{CNN}}$, SICN*, Land-sea mask
Outputs	ΔSIC	ΔSICN
Size of input data set	$2094\times17\times328\times368$	$2094\times12\times320\times360$
Size of output data set	$2094\times1\times320\times360$	$2094\times5\times320\times360$
Normalization	Inputs standardized (see main text)	Inputs standardized (see main text)
Convolution layers	4	4
Features per layer	32, 64, 128, 1	32, 64, 128, 5
Activation function(s)	ReLU, ReLU, ReLU, Linear	ReLU, ReLU, ReLU, Linear
Kernel size(s)	3×3	1×1
Kernel stride(s)	1	1
Bias parameters	False	False
Zero-padding	None	None
Total weights	98,208	11,264
Batch size	10	10
Optimizer	Adam	Adam
Learning rate	0.001	0.001
Weight decay	1×10^{-7}	1×10^{-7}
Epochs	150	125
Seed	711	711

eddies (Bolton & Zanna, 2019; Zanna & Bolton, 2020) and cloud moisture convection (Han et al., 2020). For this reason, we use them here to learn sea ice model errors, which also inherently exhibit two-dimensional structure.

4.1 Architecture

Generally speaking, a CNN can be seen as a series of linear weighted sums in which a rectangular weight matrix, or *kernel*, slides over an input image in order to produce a new feature representation of that same input. By sequentially repeating this procedure on each new feature map, and adding nonlinear activation functions between network layers, the network is then able to extract increasingly complex behaviour from the inputs, before a final operation which maps the last set of features to each pixel of the output image. Figure 3 shows this procedure in the present context of learning 'images' of sea ice DA increments. In this case we develop two independent CNNs, where each can be classified as a 'fully CNN' as the outputs of each layer are produced only by convolution operations. Network A is used to learn the aggregate (ΔSIC) increments from various atmosphere, ocean and sea ice model states and forecast tendencies, while network B uses the predictions of ΔSIC from network A in order to learn a mapping from ΔSIC to ΔSICN. We find this two-step approach yields significantly lower prediction error than using a single network to predict ΔSICN directly. Table 1 summarizes the architectural choices made for Networks A and B.

Each of the inputs of a given CNN have independent kernels that connect to every feature map in the subsequent layer of the network, hence with 3×3 kernels in each layer, 17 input variables, and features per layer of 32, 64, 128, and 1, network A has a total number of weights given by $(3\times3\times17\times32)+(3\times3\times32\times64)+(3\times3\times64\times128)+(3\times3\times128\times1)=98,208$. Meanwhile, with 1×1 kernels in each layer, 12 input variables, and features per layer of 32, 64, 128, and 5, network B has a total number of weights given by $(1\times1\times12\times32)+(1\times1\times32\times64)+(1\times1\times64\times128)+(1\times1\times128\times5)=11,264$. An advantage of the CNN approach is that a single kernel matrix is used for the entire spatial domain of a given input, meaning that structures which exhibit similar characteristics, but occur at different locations within the input, will be equally resolved. This property of *translational invariance* is not guaranteed in e.g., typical feed-forward (artificial) neural networks which use the whole domain at once as input (Gardner & Dorling, 1998). Non-linearities within the system can also be exploited by passing each feature map through a non-linear activation function, such as the rectified linear unit (ReLU) function, which is the identity function for positive values and zero for negative values. In both networks in our application, the first three convolution operations are followed by ReLU activation functions, while the final convolution to the output layer is simply linear.

The inputs to network A correspond to the 5-day means of the model states and 5-day forecast tendencies from each DA cycle, for each of SIC, sea-surface temperature (SST), zonal and meridional components of ice velocities (SIU and SIV, respectively), sea ice thickness (SIT), net shortwave radiation (SW), ice-surface skin temperature (TS), sea-surface salinity (SSS), and finally a land-sea mask containing zeros over land grid cells and ones over ocean grid cells. Note that SIU and SIV are vector fields with values located at C-grid cell edges, while the other scalar fields have values centered within each grid cell (see e.g., Griffies et al. (2004)). This means that SIU and SIV contain one additional matrix column and row, respectively, compared to the scalar fields. We therefore compute a 2-point average along the columns of SIU and rows of SIV, so that the these variables are defined on the same tracer grid as the scalar fields. The inputs to network B correspond to the ΔSIC predictions from network A, along with the model states and forecast tendencies of SICN, as well as a land-sea mask. It should also be noted that the inputs of each network (excluding the land-sea mask) are standardized by subtracting their respective mean and normalizing by their respective standard deviation, where both mean and standard deviations are computed over ocean grid cells poleward of 40° latitude, across all training samples (see section 4.2). This provides a single value of the mean and standard deviation for each network input. Furthermore, given that, in our network architecture, each convolution operation in network A reduces the size of the input image by 2 pixels in both matrix dimensions, the final outputs are 8 pixels smaller than the original inputs (hence a 9×9 centered stencil is required to make a local prediction at any grid point). To ensure we utilize the appropriate information at the image boundaries, we therefore pad the input data by 4 pixels on each side in the following way: the last 4 columns of the image are padded in front of the first column (zonal periodicity), the original first 4 columns are padded to the last column (zonal periodicity), a copy of the first 4 rows is flipped 180° counter-clockwise and padded in front of the first row (symmetry across the model's Arctic bipolar fold, see Griffies et al. (2004); the sign of the ice velocities in the first 4 rows is also flipped during this process), and finally the last row is padded with 4 rows of zeros (the final row corresponds to the Antarctic continental land mass).

4.2 Training

In order to generate accurate predictions, the weights of each CNN must be optimized. This is typically achieved by minimizing an appropriate loss function $\mathcal{L}$ which describes the similarity between the final outputs of the network and the target variable (i.e., the analysis increments). For network A the loss function ($\mathcal{L}_A$) is the mean-squared

error (MSE) of the ΔSIC predictions, while for network B the loss function ($\mathcal{L}_B$) is the sum of the MSE of each of the five ΔSICNs, as well as an additional term to impose a soft constraint that the sum of the five ΔSICNs are equal to ΔSIC:

$$\mathcal{L}_A = \frac{1}{NS}\sum_{i=1}^{NS}\left(\Delta\mathrm{SIC}_i^{\mathrm{CNN}} - \Delta\mathrm{SIC}_i^{\mathrm{True}}\right)^2, \tag{1}$$

$$\begin{aligned}\mathcal{L}_B = \sum_{k=1}^{5}\frac{1}{NS}\sum_{i=1}^{NS}\left(\Delta\mathrm{SICN}_{k_i}^{\mathrm{CNN}} - \Delta\mathrm{SICN}_{k_i}^{\mathrm{True}}\right)^2 \\ + \lambda\left(\frac{1}{NS}\sum_{i=1}^{NS}\left(\sum_{k=1}^{5}\Delta\mathrm{SICN}_{k_i}^{\mathrm{CNN}} - \sum_{k=1}^{5}\Delta\mathrm{SICN}_{k_i}^{\mathrm{True}}\right)^2\right).\end{aligned} \tag{2}$$

Here, $N = 320 \times 360 = 115,200$ is the number of model grid points, which corresponds to the entire globe. $S = 10$ is the batch size (randomly shuffled temporal samples), and $\lambda = 5$ is a scaling constant. The loss function is minimized using the Adam stochastic gradient descent method (Kingma & Ba, 2014) within the PyTorch Python library (Paszke et al., 2019), which accommodates graphical processing unit (GPU) and batch processing facilities for significant computational speed-ups and efficient memory handling, respectively. Recall Table 1 for a full list of the details of each CNN.

As well as optimizing the weights of each CNN, there are other factors which influence the predictive performance that also need to be considered. For one, there is the physical architecture of each CNN, which includes e.g., the number of layers within each network, the type of activation function, and the size of the convolution kernels. Then there are also specific hyperparameters, which include e.g., the learning rate of the Adam optimizer, and the number of training epochs. Choosing the optimal architectures and hyperparameters is referred to as *model selection* and is generally approached by selecting the model which produces the lowest error score on unseen validation data (i.e., data that were not used to optimize the CNN weights). In order to ensure that the validation error is representative of the model's predictive performance across all samples it is often necessary to perform K-fold cross-validation, where the data are split into K equal-sized temporally contiguous chunks. The model is then trained on $K-1$ chunks, and predictions are validated on the remaining chunk. We opt for temporally contiguous chunks here, as opposed to random sampling of training and validation points, due to inherent temporal auto-correlation within the data, which would likely lead to data leakage issues during the validation stage. In any case, this process is repeated K number of times where each time a different chunk is chosen to be the validation set. The average validation error across all K tests is then the generalization error of that particular CNN model. To arrive at the final CNN architectures and hyperparameters detailed in Table 1, we performed 5-fold cross-validation at each model selection step, hence for a given architecture and set of hyperparameters the model was trained 5 times, where each time the 2618 temporal samples were split into different combinations of 2094 training and 524 validation points. Specific architectures and hyperparameters were subsequently chosen based on the model which showed the lowest average 5-fold cross-validation score. Ideally, one would perform model selection by scanning all possible combinations of hyperparameters and CNN architectures and finding which combination produces the lowest cross-validation score. For large data sets however, this is computationally impractical and as such we proceeded with model selection by testing one hyperparameter and/or architecture at a time and taking the model with the lowest 5-fold cross-validation score forward to the next test (see Figure S3 for example learning curves from various model selection tests). The results in the next section are based on predictions on validation data from the final CNN models, as described in Table 1. Note that, for convenience, hereafter we refer to networks A and B together as our final network architecture.

5 Results

Before presenting the results of the CNN predictions, we first introduce the error metrics which are used to evaluate the model's performance. For a given spatial map of the SIC increments on any given day, $\Delta\text{SIC}^{\text{True}}$, and the equivalent CNN prediction on the same day, $\Delta\text{SIC}^{\text{CNN}}$, the regional uncentered spatial pattern correlation (Barnett & Schlesinger, 1987) between these two fields is given as:

$$\rho = \frac{\sum_{i=1}^{n} \Delta\text{SIC}_i^{\text{CNN}} \Delta\text{SIC}_i^{\text{True}}}{\|\Delta\text{SIC}^{\text{CNN}}\|_2 \|\Delta\text{SIC}^{\text{True}}\|_2}, \tag{3}$$

where $\|\cdot\|_2$ is the ℓ_2 vector norm, and $n = 100 \times 360 = 36,000$ for either pan-Arctic or pan-Antarctic regions (approx. 45°N and 30°S, respectively). We opt for this metric over the standard (centered) linear correlation coefficient as the subtraction of the mean to compute the covariance in the centered case may result in differences between $\Delta\text{SIC}^{\text{True}}$ and $\Delta\text{SIC}^{\text{CNN}}$ at open-ocean grid cells (e.g., Legates and Davis (1997)). Similar to the centered pattern correlation, an uncentered pattern correlation value of 1 represents a perfect agreement between the true and predicted increments on day t, while a value of -1 represents a perfect out-of-phase agreement. A value of 0 subsequently represents no agreement.

We also introduce the regional root-MSE (RMSE) as:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\Delta\text{SIC}_i^{\text{CNN}} - \Delta\text{SIC}_i^{\text{True}}\right)^2}. \tag{4}$$

This metric captures the average deviation of the predictions from the true increments, hence an RMSE value of 0 represents perfect predictions.

5.1 Predictions

In this section we show the predictions of ΔSIC as the sum of the five predicted ΔSICNs, on the held-out data that were not used to optimize the network weights during training. We therefore generate 2618 predictions spanning the 1982–2017 period, which correspond to combining the 5 individual held-out chunks from the cross-validation experiment of the final model, into a continuous time series record. We focus on ΔSIC here, as opposed to ΔSICN, as the former is the direct observable quantity and as such lends to more intuitive interpretation of the results, although the reader is referred to Figures S4–S8 for comparable versions of Figure 4 for each ΔSICN.

Figure 4 shows the seasonal climatologies of the ΔSIC predictions, where we notice that, in both hemispheres, the CNN is able to predict the average spatial pattern of the increments very well. In the Arctic, the network performs best in DJF, with average daily spatial pattern correlations of 0.73, and a spatial pattern correlation of 0.98 between the climatologies of the daily DJF predicted and true increments. The poorest predictions in the Arctic are in JJA and SON with average daily spatial pattern correlations of 0.64 and 0.62, respectively, and correlations of 0.96 and 0.98, respectively between the climatologies. In JJA for example, while the network reproduces the average spatial pattern well, the magnitude of the increments to the north of Greenland and in the Canada basin is generally too low. Similarly, in the Antarctic, the CNN also performs best in DJF with average daily spatial pattern correlations of 0.80, however the average magnitude of the predicted increments is generally too low in regions such as the Weddell Sea. The poorest predictions in the Antarctic are in MAM with average daily spatial pattern correlations of 0.64, perhaps owing to the network's inability to fully resolve the relatively small-scale heterogeneities in e.g., the Ross and Weddell seas. The large-scale patterns are generally in good accordance however. These initial results suggest that the network is able to learn the mean bias patterns of the model with considerable skill.

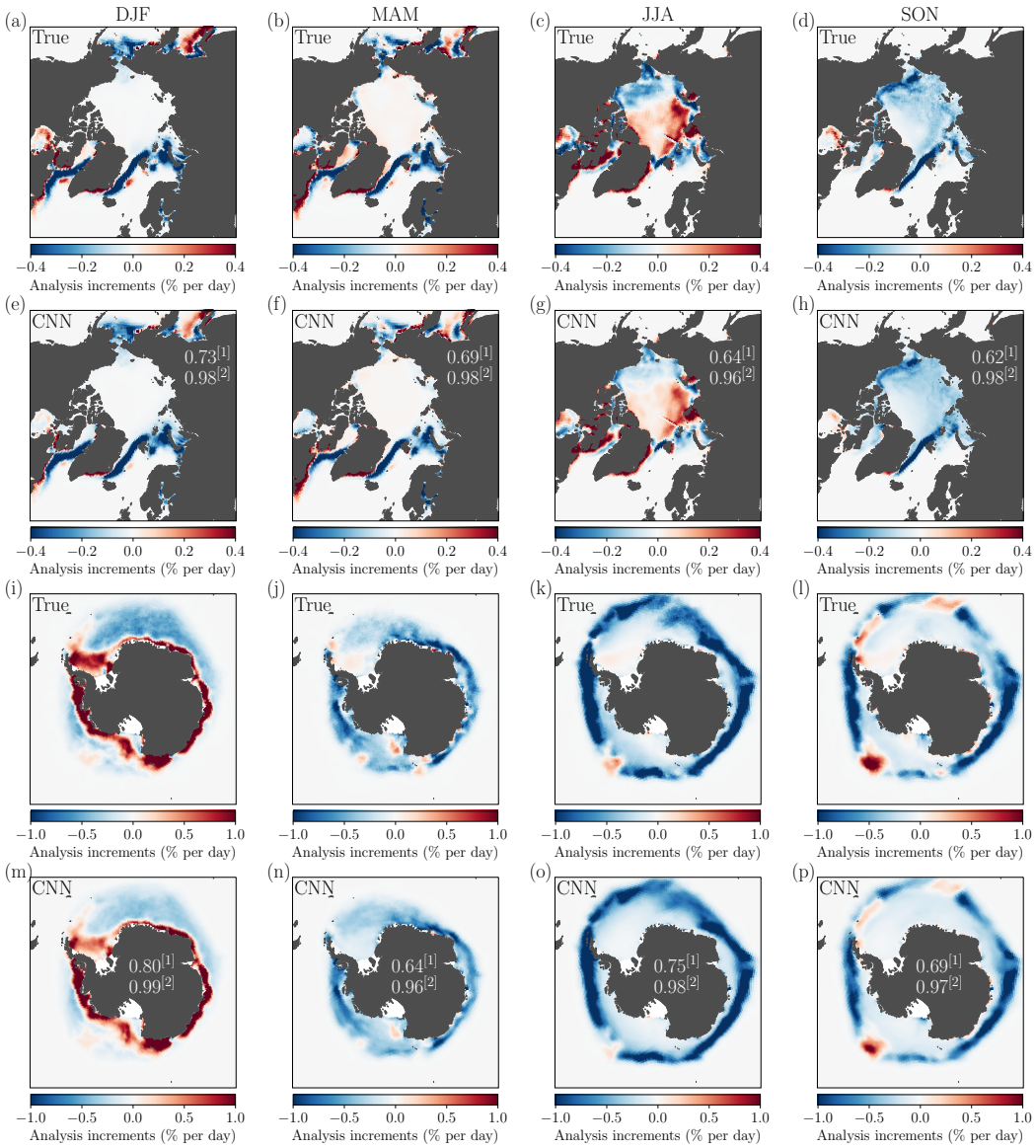

Figure 4. Seasonal climatologies of the (true) SPEAR aggregate sea ice concentration analysis increments and the equivalent CNN predictions, for both the Arctic (a)–(h) and Antarctic (i)–(p). Columns from left to right show DJF, MAM, JJA, SON climatologies, computed over the period 1982–2017. Values with the superscript [1] are the average of daily spatial pattern correlations between $\Delta\text{SIC}^{\text{True}}$ and $\Delta\text{SIC}^{\text{CNN}}$ in each respective season, while values with [2] are the spatial pattern correlations between the respective climatologies of the true and predicted increments.

Moving beyond assessments of climatologies, Figure 5 shows randomly sampled snapshots of the predictions for individual days across each season, as a way to assess how the CNN performs at capturing the fast physics errors (for an animation of the CNN performance on additional daily snapshots, see Supporting Information S2). Broadly speaking, we find that the CNN is able to capture the large-scale structure of the increments, but often fails to capture smaller-scale features. The February prediction in the Arctic (Figure 5e) shows high skill with a spatial pattern correlation of 0.74, however at this

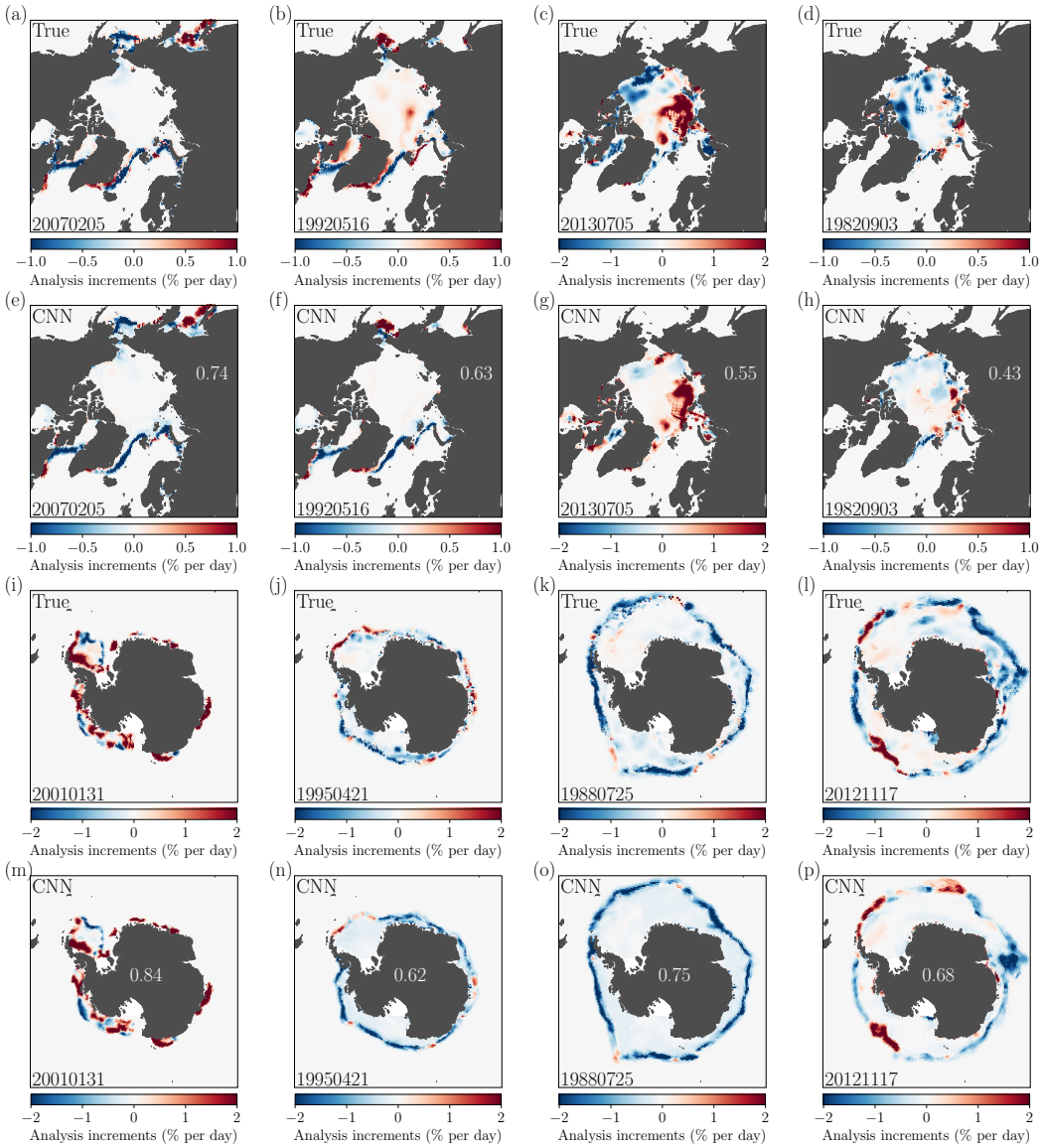

Figure 5. Daily snapshots of the (true) SPEAR aggregate sea ice concentration analysis increments and the equivalent CNN predictions, for both the Arctic (a)–(h) and Antarctic (i)–(p). Columns from left to right show random days in DJF, MAM, JJA, and SON over the period 1982–2017. Spatial pattern correlations are reported for each prediction.

time of year the increments are primarily associated with sea ice edge errors, while the increments in the central ice pack (i.e., the majority of the Arctic domain) are effectively zero. Nonetheless, the CNN is able to predict these ice edge errors very well, particularly in the Labrador, GIN, Barents, Okhotsk, and Bering seas. As the melt season progresses, the prediction skill generally drops, where it is lowest in September (Figure 5h), with a spatial pattern correlation of 0.43. The true July and September increments (Figures 5c and 5d, respectively) exhibit significant variability within the core ice pack which, in some regions, the network is unable to reproduce. For example, the large negative increments in the Beaufort and Chukchi seas in July. The CNN does however manage to

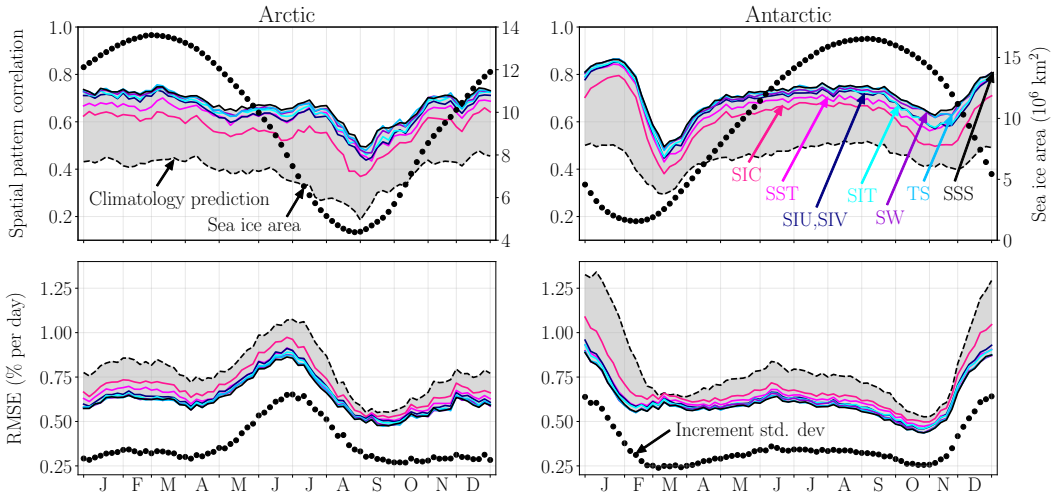

Figure 6. Prediction skill metrics for independent sensitivity tests to network inputs, presented as daily climatologies of predictions on held-out samples, computed over the period 1982–2017, for the Arctic (left column) and Antarctic (right column). The shaded region reflects the improvement in skill of the final network (solid black curve) over the benchmark climatology prediction (dashed black curve).

capture some amount of the variability in July, such as the large positive increments in the Kara and Laptev shelf seas.

The prediction skill in the Antarctic is generally higher than in the Arctic, and comparing Figures 5i and 5m, we can see that the CNN accurately predicts a significant amount of the variability in summer, with a spatial pattern correlation of 0.84. The subsequent predictions in April, July and November (Figures 5n-p) show slightly lower skill than in January, with the lowest skill in April with a spatial pattern correlation of 0.62. At these times the increments are largely related to sea ice edge errors, and the CNN is generally able to capture the large-scale patterns, as well as some of the localized features, such as the positive increments at the north-eastern edge of the Ross Sea in November (Figure 5p), and the band of positive increments along the northern edge of the Weddell Sea in April (Figure 5n).

From the daily snapshots we can infer that the CNN captures large amounts of the fast physics errors, although there is some seasonal variation to the skill, where the predictions in the Arctic are generally best over the winter period and poorest in the summer. Meanwhile in the Antarctic the predictions appear most skillful in the summer and poorest in the early growth season (April). In the next section we provide an assessment of the CNN's sensitivity to various inputs, as well as its sensitivity to the geographic training domain. In doing so, we subsequently highlight this seasonal skill variation in more detail.

5.2 Sensitivity analysis

5.2.1 Network inputs

In this section we perform sensitivity tests to determine which model states and forecast tendencies contribute most to the prediction skill of ΔSIC (again, as the sum of the five predicted ΔSICNs on held-out samples), at different times of the year. The sensitivity analysis is performed by training a series of initial networks which each con-

tain a single variable as inputs (e.g., SIC states and forecast tendencies), and assessing which of these networks results in the highest prediction skill of ΔSIC in both hemispheres. The input variable of this network is then assumed to be the most physically-relevant predictor of ΔSIC. The testing then continues by training a second series of networks which contain two input variables: the best predictor from the first test, as well as any one of the remaining input variables. The network which results in the largest improvement in skill relative to the best network from the first test is then taken forward, and so on. For 7 network input variables (classifying SIU and SIV as a single input), we therefore trained 28 independent network configurations in order to establish a hierarchy of predictors.

Figure 6 shows daily 36-year climatologies of spatial pattern correlation and RMSE error metrics, for sensitivity tests in both the Arctic and Antarctic domains. The hierarchy of predictors in terms of largest skill contribution proceeds as: SIC, SST, SIU and SIV, SIT, SW, TS, and finally SSS. Hence for the SIC curves, the network inputs to generate these predictions are only SIC states and forecast tendencies, while for the SST curves, the network inputs are SIC and SST states and forecast tendencies, and so on. The SSS curve then represents the predictions from the final model (i.e., the network architecture presented in section 4.1). The climatology prediction (black dashed curve) refers to using the daily 36-year climatology of the true ΔSIC increments to predict the true ΔSIC increment on any given day. This is an offline-equivalent to the 'ocean tendency adjustment' approach by Lu et al. (2020), as discussed in section 1, and as such serves as our benchmark here, where we can see that each sensitivity test provides improvement in skill over this climatological tendency benchmark. From this analysis we can also see that, relative to the benchmark climatology, SIC is responsible for a significant fraction of the overall network skill (approx. 66% in both hemispheres). SST, SIU and SIV then account for an additional 20%, with the remaining variables SIT, SW, TS and SSS making up the last 14%. Furthermore, while SIC, SST, SIU and SIV are essential inputs in all months of the year, the contributions from other variables such as SW and TS are generally limited to the summer months.

In terms of spatial pattern correlation, the maximum skill of the final network in the Arctic occurs at the beginning of March, after which the skill declines somewhat continuously until the end of July, and then more rapidly to its minimum in early September. Meanwhile in the Antarctic, the points of maximum and minimum skill are separated by approximately 1.5 months, with the maximum occurring at the end of January, and the minimum at the beginning of March. Although the skill variation in the Arctic appears to somewhat correlate with the climatological sea ice area, the rate of change in sea ice area in the melt and growth season is generally not consistent with that of the CNN prediction skill. Furthermore, in the Antarctic the skill is increasing between November and January, while the sea ice area is decreasing. This therefore suggests that the skill variation is not directly tied to the seasonal cycle of sea ice area. When also considering the standard deviation of the increments, we can see that the low spatial pattern correlation scores coincide with times when the standard deviation of the increments, and hence the RMSE, are relatively low. This may initially suggest that the lower spatial pattern correlation at these times is either a consequence of low signal variance, or that the network training does little to optimize these points as they inherently have lower MSE than e.g., the winter months. If however the spatial pattern correlation scores were a direct reflection of the increment standard deviation, we would expect to see similarly low spatial pattern correlations in e.g., April in the Arctic or November in the Antarctic, however this is not generally the case. What is noticeable, is that the climatology benchmark also exhibits the same seasonal variation in spatial pattern correlation and RMSE as the CNN predictions, highlighting that the lower skill in the late summer in both hemispheres is not likely due to any shortcomings in the ML model, but rather a feature of the increments themselves. In particular, the climatological prediction is less

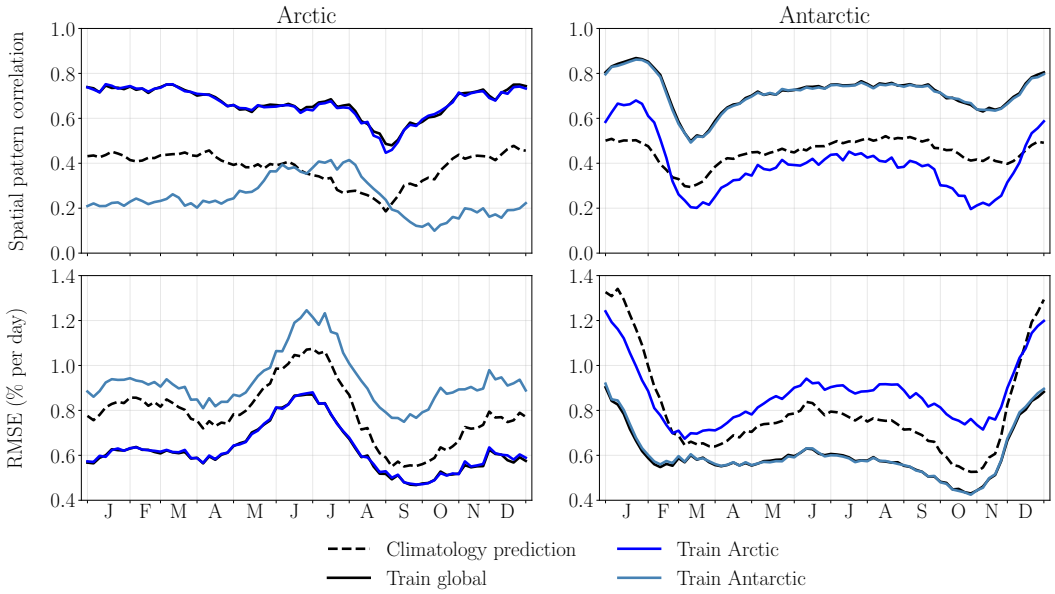

Figure 7. Prediction skill metrics for independent sensitivity tests to the network training domain, presented as daily climatologies of predictions on held-out samples, computed over the period 1982–2017, for the Arctic (left column) and Antarctic (right column).

skillful in the low-CNN-skill months, suggesting that these months are inherently more challenging to predict.

5.2.2 Training domain

The network in this study is trained on data from the entire globe, meaning that it must find the optimal set of weights which generalize to make accurate predictions of the analysis increments in both the Arctic and the Antarctic. Given that the bias patterns, and hence characteristics of the increments, are somewhat different between the two hemispheres, we conduct further sensitivity tests to determine how well the network has generalized. As before, error metrics are shown in terms of the sum of the five predicted ΔSICNs on held-out samples that were not used to train the model.

Figure 7 shows daily climatologies of spatial pattern correlation and RMSE error metrics, for three variations of the network training setup. One where the network is trained on the entire globe (i.e., our proposed network in section 4.1), one where the network is trained on just the Arctic domain, and one where the network is trained on just the Antarctic domain. Here we notice that the network which is trained on global data is able to make just as skillful predictions of ΔSIC in the Arctic, as the network which is trained only on Arctic data. The same is also true for the Antarctic case. Interestingly, we can also see that the network which is trained only on the Arctic are still able to make relatively skillful predictions of the Antarctic increments, even performing better than the benchmark climatology predictions between the months of December and February. Meanwhile, the network which is trained on Antarctic data is not able to generalize as well to the Arctic, although still shows some small amount of skill between July and August. This analysis therefore confirms that training on global data is vital for generalizing across domains while still matching the skill of networks trained on each individual domain.

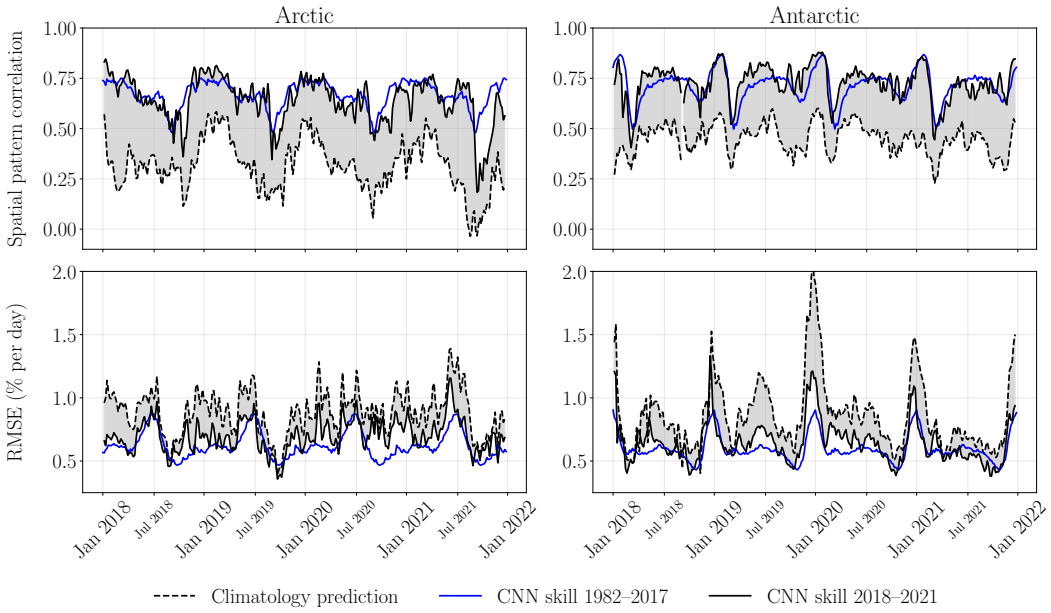

Figure 8. Generalization performance of CNN predictions for the extended period between January 2018 and December 2021. Error metrics for the black curves are shown at the frequency of the data assimilation system (5-daily), while the blue curve is the daily climatology skill of the final network over the 1982–2017 period.

5.3 Final validation

Due to the fact that we perform model selection by choosing specific CNN architectures and hyperparameters which minimize the average cross-validation score on data that were not used to optimize the CNN weights, there is an inherent risk of over-fitting the model to these validation data. As such, it is often necessary to retain an additional data set which has not been used for validation at any point during the model selection process. For this, we extend the Y. Zhang et al. (2021) sea ice DA experiment from December 27$^{\text{th}}$ 2017, through to December 27$^{\text{th}}$ 2021, providing an additional 291 validation data points. We subsequently evaluate the performance of our CNN model by training on all 2618 samples between 1982–2017, and validating on the extended data period between 2018–2021. It should be noted that this extended DA experiment is identical in configuration to that which was outlined in section 2.2, except that in this extended case the atmospheric forcing from JRA55-do reanalysis corresponds to version 1.5, while previously it was version 1.3. This version change relates to a correction in the sign and rotation of tropical cyclones, and as such we do not expect this to result in significant differences in the representation of sea ice in the extended DA simulations.

Figure 8 shows daily spatial pattern correlation and RMSE error metrics over the 2018–2021 period for both the Arctic and Antarctic domains (black curves). We also overlay the daily climatology skill of the final network architecture from the cross-validation experiments between 1982–2017, hence the blue curves here are identical to the 'Train global' curves in Figure 7, and are simply repeated for each of the 4 validation years presented. The predictions appear to generalize well to the future data, where spatial pattern correlation values are generally in accordance with the 1982–2017 period, particularly in the Antarctic, and are still out-performing the climatology prediction in both hemispheres. On average, the RMSE over the 2018–2021 period is slightly higher than the 1982–2017 climatology, which is due to the fact that there is a non-stationary com-

ponent to the increments, whereby the variance increases over the course of the time series record (see Figure S9). Therefore naturally the climatological RMSE of the CNN predictions increases over time as well (see Figure S10). In any case, the generalization ability of the predictions suggests that the CNN has not simply over-fitted to the training and/or validation data during model selection.

6 Discussion

The ability of the proposed CNN to make skillful predictions of the sea ice concentration analysis increments, using only information on local model state variables and their tendencies, provides interesting avenues for future work. The fact that the predictions show improvements in skill relative to a daily increment climatology (e.g., Lu et al. (2020)), generalize well to each hemisphere, and show skill on a separate validation data set, strongly suggests that the CNN could be used to reduce sea ice biases within SPEAR, either as an online sea ice model parameterization, or as a bias correction tool for numerical sea ice prediction. Ultimately, one could argue that there is still room for improvement in the CNN performance, particularly in the late summer months. Considering the inherent complexity of the problem at hand, and the likely influence of both non-linear and non-local processes, it is conceivable to push the limit of predictive skill further by increasing the complexity of the network, both in terms of the total number of weights, and the 9×9 grid cell domain of influence on a local prediction. Indeed, such changes could be implemented through increasing the width and/or depth of the network, as well as incorporating non-local connections (in space) through e.g., fully-connected layers. On the other hand, the architectures here been developed specifically with the goal of a sea ice model parameterization in mind, and as such, factors including computational cost and practicality of implementation in parallelized high-performance computing environments have been considered throughout the development. In the following sections we provide a discussion on the directions for future work relating to both sea ice parameterization and seasonal sea ice prediction.

6.1 Considerations for parameterization

ML models have been shown to be successful at parameterizing subgrid-scale processes within dynamical models, including ocean mesoscale eddies (Guillaumin & Zanna, 2021), atmospheric convection (Yuval & O'Gorman, 2020), and sea ice dynamics (Finn et al., 2023). Common to each of these studies is that the ML models target specific physical processes, with the aim of replacing pre-existing knowledge-based parameterizations, or deriving new parameterizations for physical processes which are not currently represented. On the other hand, our proposed CNN is trained to predict sea ice increments which reflect numerous interacting model errors across various model components. To subsequently disentangle these coupled model physics errors a-posteriori and then apply them as parameterizations to their respective components, is not straightforward. In our goal of constructing a sea ice model parameterization, it is critical to ensure that the parameterization is not acting to correct coupled model errors that originate in other model components (e.g., an ocean heat transport bias or atmospheric circulation bias that imprints upon the sea ice). Our DA-ML methodology attempts to mitigate this possibility, as the ice-ocean DA system is driven by atmospheric reanalysis and also nudges SST and SSS towards observed values. These observational constraints on the atmosphere and ocean allow us to interpret the DA increments as isolated sea ice model physics errors, however this assumption is not perfect as the ocean component of the DA system can still imprint some errors on the sea ice state (e.g., Figure S1). Future investigation will be required to determine how the CNN generalizes in a fully coupled setting with fully-interactive atmosphere-ice-ocean feedbacks (see section 6.2).

Another major consideration for a sea ice parameterization is how to appropriately conserve mass, heat, and salt. In the context of the ocean, Lu et al. (2020) achieved global conservation of heat and salinity when implementing the climatological ocean DA increments into MOM6 by ensuring that the global integral of the correction to each variable was zero. In the case of sea ice, assuming the parameterization enters the thermodynamic solver, then appropriately coupling this parameterization with the upper ocean would mean that a predicted negative sea ice concentration increment would remove sea ice (column-wise) by adding mass and salt to the ocean mixed layer, while also removing heat. This step is likely to come in the form of a mass, heat, and salt budget assessment between the ice and ocean after evaluating the amount of local sea ice mass change associated with a given predicted SIC increment, rather than adapting the CNN architectures themselves to respect conservation.

Although we have considered implementation cost in the design of our network, some investigation will be required to quantify this cost in terms of both matrix computations and additional memory load. Regarding memory load, our parameterization will not require any additional memory in terms of the number of grid cells stored on any one central processing unit (CPU), as our 9×9 network stencil requires the same number of 'halo' grid points as the default SPEAR configuration, which uses a halo size of 4. There will be some small amount of memory cost for storing the network weights on each CPU however. Looking to similar studies, Guillaumin and Zanna (2021) found that implementing a fully CNN with 8 convolutional layers as a stochastic parameterization into an idealized shallow water model resulted in a 25% increase in the run time, compared to an unparameterized simulation. C. Zhang et al. (2023) also found that the cost of doing inference with this same network as a parameterization in MOM6 was 10 times more expensive than the CPU cost of the simulation itself. Although we effectively have 8 convolutional layers when considering both networks A and B, we can still expect much lower computational overheads given that ours is a deterministic model (i.e., we predict a single output at each grid point for each ΔSICN, rather a, potentially larger, number of parameters which describe a distribution of values), and that our kernel size for network B is 1×1 in each layer, while the Guillaumin and Zanna (2021) network uses variable size kernels throughout, ranging from sizes 3×3 to 5×5. Like in this study, they also did not use zero-padding, though in their case given the larger kernel sizes, they required a stencil of 21×21 grid points to make a local prediction.

Finally, the increments in this study represent error growth over a 5-day period, and the input states and tendencies of the CNN are given as 5-day means. After implementation, the CNN predictions will need to produce a correction which reflects error growth over a given model timestep, and similarly the input states and tendencies will need to be adjusted accordingly. This will therefore require further sensitivity tests to determine how to appropriately perform this scaling.

6.2 Considerations for sea ice forecasting

Some of the initial concerns over implementation of the CNN as a sea ice model parameterization can be alleviated by assessing how the network performs as an online bias correction tool within the context of seasonal sea ice forecasting. In previous work, Y. Zhang et al. (2022) showcased the benefits of using SIC assimilation to initialize the sea ice conditions for SPEAR retrospective forecasts (hereafter reforecasts) of the Arctic sea ice cover between 1992–2017. In Y. Zhang et al. (2022), the same ice-ocean SPEAR model configuration and initial conditions as outlined here in section 2.2, were used to perform DA between 1982 and the first day of each month, for all years between 1992 and 2017. Whereby the first day of each month represented the initialization point, after which the model would run in fully coupled mode to generate forecasts out to 1-year lead time. Assimilation in Y. Zhang et al. (2021, 2022) was performed by passing the prior model state variables and observations to the Data Assimilation Research Testbed

(DART; Anderson et al. (2009)), which then computes the set of analysis states offline, providing the new set of initial conditions with which to begin the next assimilation cycle. Given that our CNN is inherently independent of the observations, we propose that it would be relatively straightforward to bias correct the sea ice within the fully coupled reforecast period by replacing the standard call to DART with our CNN. In this scenario, we could perform seasonal reforecasts (up to 12-month lead times) to assess how the network generalizes to the fully coupled SPEAR model, while not requiring strict conservation properties due to the shorter time scales. Furthermore, we could continue in the same 5-day cycle configuration so that the network predictions would not need to be scaled for different temporal sampling. If the reforecasts then have improved skill relative to the SPEAR DA-initialized reforecasts from Y. Zhang et al. (2022), they may be fit-for-purpose as a model parameterization.

7 Concluding remarks

7.1 Summary

In this study we have shown that deep learning (DL), specifically convolutional neural networks (CNNs), can be used to make skillful predictions of sea ice model errors, in the form of data assimilation (DA) increments, using only information from model state variables and tendencies (the time derivative of the model state variables). We developed a CNN using an ice-ocean DA system which assimilates satellite observations of sea ice concentration (SIC) into the Seamless system for Prediction and EArth system Research (SPEAR) model every 5 days between 1982–2017. SPEAR has a 5-category ice thickness distribution, hence concentration increments are produced for each subgrid category, where the observable (aggregate) SIC increment corresponds to the sum of 5 categories. We therefore developed a two-step CNN architecture, in which the first step learns the physical mapping from various local sea ice, ocean and atmosphere state variables and forecast tendencies to the aggregate SIC increments. The second step then learns the mapping from the aggregate concentration error to each of the subgrid terms. We subsequently showed that our DL architecture is able to make skillful predictions of the SIC increments in both the Arctic and the Antarctic and across all seasons. Spatial pattern correlations between the climatologies of the observed and predicted increments are high, with values of at least 0.96 for both the Arctic and Antarctic, demonstrating that the CNN is able to skillfully capture the mean model bias. The CNN also has skill at predicting the state-dependent model errors, with daily pattern correlation values ranging from 0.64–0.80 and 0.62–0.73 in the Antarctic and Arctic, respectively. This shows that the CNN is able to predict the fast physics errors and systematic bias patterns of the SPEAR model with considerable skill, which is also confirmed by the fact that the predictions show improved skill over a model which simply predicts the climatological mean increment on any given day of the year. Sensitivity analysis revealed that SIC as an input to the network is responsible for approximately 66% of the overall network skill, followed by sea-surface temperature (SST) and ice velocities which account for 20%, and finally ice thickness, net shortwave radiation, ice-surface skin temperature and sea-surface salinity which account for the remaining 14%.

7.2 Outlook

Recent studies have highlighted how DA provides a unique opportunity to leverage sparse and/or noisy observations, in order to facilitate machine learning of structural model errors (Bonavita & Laloyaux, 2020; Brajard et al., 2021; Farchi et al., 2021; Mojgani et al., 2022; Chen et al., 2022). Building on this, we have shown here how DA also provides the ability to learn errors within unobserved model state variables, and hence provides a new framework for learning subgrid-scale parameterizations for climate models. In section 6 we subsequently outlined how the strong predictive performance of the

CNN and its generalization ability suggests that the network has the potential to reduce sea ice biases in free-running climate simulations, as a sea ice model parameterization within SPEAR. Irrespective of this eventual goal however, the findings in this work ultimately have wider implications for the climate modeling and numerical weather prediction (NWP) community in general. With regards to NWP, previous studies have already shown that ML techniques can be used to learn state-dependent fast physics errors within large-scale atmospheric models, subsequently leading to improved online predictions by using the ML model as a bias correction tool (Bonavita & Laloyaux, 2020; Chen et al., 2022). In our study, we have shown that the concept of learning state-dependent fast physics errors is transferable to a global ice-ocean model, which could further aid NWP when considering that coupling the atmosphere with an ice-ocean model has previously shown to improve short-term weather predictions (G. Smith et al., 2018).

Turning to longer-term simulations, the fact that the systematic errors are also predictable suggests that a parameterization built from DA increments has the potential to reduce persistent climate model biases and improve the fidelity of climate change projections. On the other hand, while we have shown that state variables such as SIC and SST explain a significant fraction of the variance in the analysis increments, our current framework does not allow us to attribute these correlations to a specific model deficiency, for example an incorrectly parameterized or missing physical process. One additional avenue for future work could therefore involve designing a perfect model experiment in which a single ensemble member is run with a specific parameterization that has been tuned or turned on (e.g., sea ice ridging or melt-pond formation). This member would then be treated as the ground truth and assimilated into the original model. The resultant analysis increments would then be a manifestation of the systematic bias within the original model, associated with this specific incorrect/missing parameterization, and hence one could more confidently isolate which state variables within an ML model contribute most to predicting this particular structural error.

8 Open Research

All data for training each CNN are openly available at the following locations:

- Inputs (DA forecast states and tendencies): `ftp://sftp.gfdl.noaa.gov/perm/William.Gregory/seaice_DA-ML_inputs_1982-2017.nc`
- Outputs (DA increments): `ftp://sftp.gfdl.noaa.gov/perm/William.Gregory/seaice_DA-ML_outputs_1982-2017.nc`

Python code to pre-process the input data and train the CNNs can also be found at `https://github.com/m2lines/seaice_DA-ML`. The optimized weights of the CNNs and standardization statistics for the inputs are also saved within the same repository.

Acknowledgments

William Gregory, Mitchell Bushuk, Alistair Adcroft and Laure Zanna received M^{2}LInES research funding by the generosity of Eric and Wendy Schmidt by recommendation of the Schmidt Futures program. This work was also intellectually supported by various other members of the M^{2}LInES project, as well as being supported through the provisions of computational resources from the National Oceanic and Atmospheric Administration (NOAA) Geophysical Fluid Dynamics Laboratory (GFDL). We also thank Spencer Clark and Zachary Labe for their invaluable feedback on this article.

References

Adcroft, A., Anderson, W., Balaji, V., Blanton, C., Bushuk, M., Dufour, C. O., ...

others (2019). The GFDL global ocean and sea ice model OM4.0: Model description and simulation features. *Journal of Advances in Modeling Earth Systems*, *11*(10), 3167–3211. doi: https://doi.org/10.1029/2019MS001726

Anderson, J. (2001). An ensemble adjustment Kalman filter for data assimilation. *Monthly weather review*, *129*(12), 2884–2903. doi: https://doi.org/10.1175/1520-0493(2001)129⟨2884:AEAKFF⟩2.0.CO;2

Anderson, J., Hoar, T., Raeder, K., Liu, H., Collins, N., Torn, R., & Avellano, A. (2009). The data assimilation research testbed: A community facility. *Bulletin of the American Meteorological Society*, *90*(9), 1283–1296. doi: https://doi.org/10.1175/2009BAMS2618.1

Auligné, T., McNally, A., & Dee, D. (2007). Adaptive bias correction for satellite data in a numerical weather prediction system. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, *133*(624), 631–642. doi: https://doi.org/10.1002/qj.56

Balmaseda, M., Dee, D., Vidard, A., & Anderson, D. (2007). A multivariate treatment of bias for sequential data assimilation: Application to the tropical oceans. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, *133*(622), 167–179. doi: https://doi.org/10.1002/qj.12

Banzon, V., Smith, T. M., Chin, T. M., Liu, C., & Hankins, W. (2016). A long-term record of blended satellite and in situ sea-surface temperature for climate monitoring, modeling and environmental studies. *Earth System Science Data*, *8*(1), 165–176. doi: https://doi.org/10.5194/essd-8-165-2016

Barnett, T. P., & Schlesinger, M. E. (1987). Detecting changes in global climate induced by greenhouse gases. *Journal of Geophysical Research: Atmospheres*, *92*(D12), 14772–14780. doi: https://doi.org/10.1029/JD092iD12p14772

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2022). Pangu-Weather: A 3D high-resolution model for fast and accurate global weather forecast. *arXiv preprint arXiv:2211.02556*. doi: https://doi.org/10.48550/arXiv.2211.02556

Bitz, C. M., Holland, M. M., Weaver, A. J., & Eby, M. (2001). Simulating the ice-thickness distribution in a coupled climate model. *Journal of Geophysical Research: Oceans*, *106*(C2), 2441–2463. doi: https://doi.org/10.1029/1999JC000113

Bitz, C. M., & Lipscomb, W. H. (1999). An energy-conserving thermodynamic model of sea ice. *Journal of Geophysical Research: Oceans*, *104*(C7), 15669–15677. doi: https://doi.org/10.1029/1999JC900100

Bolton, T., & Zanna, L. (2019). Applications of deep learning to ocean data inference and subgrid parameterization. *Journal of Advances in Modeling Earth Systems*, *11*(1), 376–399. doi: https://doi.org/10.1029/2018MS001472

Bonavita, M., & Laloyaux, P. (2020). Machine learning for model error inference and correction. *Journal of Advances in Modeling Earth Systems*, *12*(12), e2020MS002232. doi: https://doi.org/10.1029/2020MS002232

Bouillon, S., Maqueda, M. A. M., Legat, V., & Fichefet, T. (2009). An elastic–viscous–plastic sea ice model formulated on Arakawa b and c grids. *Ocean Modelling*, *27*(3-4), 174–184. doi: https://doi.org/10.1016/j.ocemod.2009.01.004

Brajard, J., Carrassi, A., Bocquet, M., & Bertino, L. (2021). Combining data assimilation and machine learning to infer unresolved scale parametrization. *Philosophical Transactions of the Royal Society A*, *379*(2194), 20200086. doi: https://doi.org/10.1098/rsta.2020.0086

Brenowitz, N. D., & Bretherton, C. S. (2018). Prognostic validation of a neural network unified physics parameterization. *Geophysical Research Letters*, *45*(12), 6289–6298. doi: https://doi.org/10.1029/2018GL078510

Bretherton, C. S., Henn, B., Kwa, A., Brenowitz, N. D., Watt-Meyer, O., McGib-

bon, J., ... Harris, L. (2022). Correcting coarse-grid weather and climate models by machine learning from global storm-resolving simulations. *Journal of Advances in Modeling Earth Systems*, *14*(2), e2021MS002794. doi: https://doi.org/10.1029/2021MS002794

Briegleb, B., & Light, B. (2007). A delta-eddington mutiple scattering parameterization for solar radiation in the sea ice component of the Community Climate System Model. *University Corporation for Atmospheric Research*. doi: https://doi.org/10.5065/D6B27S71

Carrassi, A., & Vannitsem, S. (2011). Treatment of the error due to unresolved scales in sequential data assimilation. *International Journal of Bifurcation and Chaos*, *21*(12), 3619–3626. doi: https://doi.org/10.1142/S0218127411030775

Cavalieri, D. J., Parkinson, C. L., Gloersen, P., & Zwally, H. J. (1996). Sea ice concentrations from Nimbus-7 SMMR and DMSP SSM/I-SSMIS passive microwave data, version 1. NASA Natl. Snow and Ice Data Cent. *Distrib. Active Arch. Cent., Boulder, Colo.*. doi: https://doi.org/10.5067/8GQ8LZQVL0VL

Chen, T.-C., Penny, S. G., Whitaker, J. S., Frolov, S., Pincus, R., & Tulich, S. (2022). Correcting systematic and state-dependent errors in the NOAA FV3-GFS using neural networks. *Journal of Advances in Modeling Earth Systems*, *14*(11). doi: https://doi.org/10.1029/2022MS003309

Chepurin, G. A., Carton, J. A., & Dee, D. (2005). Forecast model bias correction in ocean data assimilation. *Monthly weather review*, *133*(5), 1328–1342. doi: https://doi.org/10.1175/MWR2920.1

Crawford, W., Frolov, S., McLay, J., Reynolds, C. A., Barton, N., Ruston, B., & Bishop, C. H. (2020). Using analysis corrections to address model error in atmospheric forecasts. *Monthly Weather Review*, *148*(9), 3729–3745. doi: https://doi.org/10.1175/MWR-D-20-0008.1

Dansereau, V., Weiss, J., Saramito, P., & Lattes, P. (2016). A maxwell elasto-brittle rheology for sea ice modelling. *The Cryosphere*, *10*(3), 1339–1359. doi: https://doi.org/10.5194/tc-10-1339-2016

Dawson, G., Landy, J., Tsamados, M., Komarov, A. S., Howell, S., Heorton, H., & Krumpen, T. (2022). A 10-year record of Arctic summer sea ice freeboard from CryoSat-2. *Remote Sensing of Environment*, *268*, 112744. doi: https://doi.org/10.1016/j.rse.2021.112744

Dee, D. P. (2004). Variational bias correction of radiance data in the ECMWF system. In *Proceedings of the ecmwf workshop on assimilation of high spectral resolution sounders in nwp, reading, uk* (Vol. 28, pp. 97–112).

Dee, D. P. (2005). Bias and data assimilation. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, *131*(613), 3323–3343. doi: https://doi.org/10.1256/qj.05.137

Dee, D. P., & Uppala, S. (2009). Variational bias correction of satellite radiance data in the ERA-Interim reanalysis. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, *135*(644), 1830–1841. doi: https://doi.org/10.1002/qj.493

Delworth, T. L., Cooke, W. F., Adcroft, A., Bushuk, M., Chen, J.-H., Dunne, K. A., ... others (2020). SPEAR: The next generation GFDL modeling system for seasonal to multidecadal prediction and projection. *Journal of Advances in Modeling Earth Systems*, *12*(3), e2019MS001895. doi: https://doi.org/10.1029/2019MS001895

Dong, C., Loy, C. C., He, K., & Tang, X. (2015). Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, *38*(2), 295–307. doi: https://doi.org/10.1109/TPAMI.2015.2439281

Farchi, A., Laloyaux, P., Bonavita, M., & Bocquet, M. (2021). Using machine learning to correct model error in data assimilation and forecast applications.

Quarterly Journal of the Royal Meteorological Society, *147*(739), 3067–3084. doi: https://doi.org/10.1002/qj.4116

Finn, T., Durand, C., Farchi, A., Bocquet, M., Chen, Y., Carrassi, A., & Dansereau, V. (2023). Deep learning of subgrid-scale parametrisations for short-term forecasting of sea-ice dynamics with a maxwell-elasto-brittle rheology. *EGUsphere*, 2022-1342. doi: https://doi.org/10.5194/egusphere-2022-1342

Flocco, D., Schroeder, D., Feltham, D. L., & Hunke, E. C. (2012). Impact of melt ponds on Arctic sea ice simulations from 1990 to 2007. *Journal of Geophysical Research: Oceans*, *117*(C9). doi: https://doi.org/10.1029/2012JC008195

Gardner, M. W., & Dorling, S. (1998). Artificial neural networks (the multilayer perceptron) — a review of applications in the atmospheric sciences. *Atmospheric environment*, *32*(14-15), 2627–2636. doi: https://doi.org/10.1016/S1352-2310(97)00447-0

Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G., & Yacalis, G. (2018). Could machine learning break the convection parameterization deadlock? *Geophysical Research Letters*, *45*(11), 5742–5751. doi: https://doi.org/10.1029/2018GL078202

Griffies, S. M., Harrison, M. J., Pacanowski, R. C., & Rosati, A. (2004). A technical guide to MOM4. *GFDL Ocean Group Tech. Rep*, *5*, 342.

Grundner, A., Beucler, T., Gentine, P., Iglesias-Suarez, F., Giorgetta, M. A., & Eyring, V. (2022). Deep learning based cloud cover parameterization for ICON. *Journal of Advances in Modeling Earth Systems*, e2021MS002959. doi: https://doi.org/10.1029/2021MS002959

Guillaumin, A. P., & Zanna, L. (2021). Stochastic-deep learning parameterization of ocean momentum forcing. *Journal of Advances in Modeling Earth Systems*, *13*(9), e2021MS002534. doi: https://doi.org/10.1029/2021MS002534

Han, Y., Zhang, G. J., Huang, X., & Wang, Y. (2020). A moist physics parameterization based on deep learning. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2020MS002076. doi: https://doi.org/10.1029/2020MS002076

Hibler, W. D. (1979). A dynamic thermodynamic sea ice model. *Journal of physical oceanography*, *9*(4), 815–846. doi: https://doi.org/10.1175/1520-0485(1979)009⟨0815:ADTSIM⟩2.0.CO;2

Horvat, C., & Tziperman, E. (2015). A prognostic model of the sea-ice floe size and thickness distribution. *The Cryosphere*, *9*(6), 2119–2134. doi: https://doi.org/10.5194/tc-9-2119-2015

Kern, S., Lavergne, T., Notz, D., Pedersen, L. T., & Tonboe, R. (2020). Satellite passive microwave sea-ice concentration data set inter-comparison for Arctic summer conditions. *The Cryosphere*, *14*(7), 2469–2493. doi: https://doi.org/10.5194/tc-14-2469-2020

Kern, S., Lavergne, T., Notz, D., Pedersen, L. T., Tonboe, R. T., Saldo, R., & Sørensen, A. M. (2019). Satellite passive microwave sea-ice concentration data set intercomparison: closed ice and ship-based observations. *The Cryosphere*, *13*(12), 3261–3307. doi: https://doi.org/10.5194/tc-13-3261-2019

Kim, H., Kang, S. M., Kay, J. E., & Xie, S.-P. (2022). Subtropical clouds key to Southern Ocean teleconnections to the tropical Pacific. *Proceedings of the National Academy of Sciences*, *119*(34), e2200514119. doi: https://doi.org/10.1073/pnas.2200514119

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. doi: https://doi.org/10.48550/arXiv.1412.6980

Klinker, E., & Sardeshmukh, P. D. (1992). The diagnosis of mechanical dissipation in the atmosphere from large-scale balance requirements. *Journal of Atmospheric Sciences*, *49*(7), 608–627. doi: https://doi.org/10.1175/1520-0469(1992)049⟨0608:TDOMDI⟩2.0.CO;2

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–

90. doi: https://doi.org/10.1145/3065386

Laloyaux, P., Bonavita, M., Dahoui, M., Farnan, J., Healy, S., Hólm, E., & Lang, S. (2020). Towards an unbiased stratospheric analysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 2392–2409. doi: https://doi.org/10.1002/qj.3798

Laloyaux, P., Kurth, T., Dueben, P. D., & Hall, D. (2022). Deep learning to estimate model biases in an operational NWP assimilation system. *Journal of Advances in Modeling Earth Systems*, e2022MS003016. doi: https://doi.org/10.1029/2022MS003016

Landy, J. C., Dawson, G. J., Tsamados, M., Bushuk, M., Stroeve, J. C., Howell, S. E., ... others (2022). A year-round satellite sea-ice thickness record from CryoSat-2. *Nature*, 1–6. doi: https://doi.org/10.1038/s41586-022-05058-5

Large, W., & Danabasoglu, G. (2006). Attribution and impacts of upper-ocean biases in CCSM3. *Journal of Climate*, *19*(11), 2325–2346. doi: https://doi.org/10.1175/JCLI3740.1

Legates, D. R., & Davis, R. E. (1997). The continuing search for an anthropogenic climate change signal: Limitations of correlation-based approaches. *Geophysical Research Letters*, *24*(18), 2319–2322. doi: https://doi.org/10.1029/97GL02207

Leith, C. (1978). Objective methods for weather prediction. *Annual Review of Fluid Mechanics*, *10*(1), 107–128. doi: https://doi.org/10.1146/annurev.fl.10.010178.000543

Lu, F., Harrison, M. J., Rosati, A., Delworth, T. L., Yang, X., Cooke, W. F., ... others (2020). GFDL's SPEAR seasonal prediction system: Initialization and ocean tendency adjustment (OTA) for coupled model predictions. *Journal of Advances in Modeling Earth Systems*, *12*(12), e2020MS002149. doi: https://doi.org/10.1029/2020MS002149

Mitchell, L., & Carrassi, A. (2015). Accounting for model error due to unresolved scales within ensemble kalman filtering. *Quarterly Journal of the Royal Meteorological Society*, *141*(689), 1417–1428. doi: https://doi.org/10.1002/qj.2451

Mojgani, R., Chattopadhyay, A., & Hassanzadeh, P. (2022). Discovery of interpretable structural model errors by combining Bayesian sparse regression and data assimilation: A chaotic Kuramoto–Sivashinsky test case. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *32*(6), 061105. doi: https://doi.org/10.1063/5.0091282

Murphy, J. M., Sexton, D. M., Barnett, D. N., Jones, G. S., Webb, M. J., Collins, M., & Stainforth, D. A. (2004). Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, *430*(7001), 768–772. doi: https://doi.org/10.1038/nature02771

Murphy, K. (2022). *Probabilistic machine learning: An introduction.* Cambridge MIT press.

Notz, D. (2012). Challenges in simulating sea ice in Earth system models. *Wiley Interdisciplinary Reviews: Climate Change*, *3*(6), 509–526. doi: https://doi.org/10.1002/wcc.189

O'Gorman, P. A., & Dwyer, J. G. (2018). Using machine learning to parameterize moist convection: Potential for modeling of climate, climate change, and extreme events. *Journal of Advances in Modeling Earth Systems*, *10*(10), 2548–2563. doi: https://doi.org/10.1029/2018MS001351

Ólason, E., Boutin, G., Korosov, A., Rampal, P., Williams, T., Kimmritz, M., ... Samaké, A. (2022). A new brittle rheology and numerical framework for large-scale sea-ice models. *Journal of Advances in Modeling Earth Systems*, *14*(8), e2021MS002685. doi: https://doi.org/10.1029/2021MS002685

Palmer, T., & Weisheimer, A. (2011). Diagnosing the causes of bias in climate models – why is it so hard? *Geophysical & Astrophysical Fluid Dynamics*, *105*(2-3), 351–365. doi: https://doi.org/10.1080/03091929.2010.547194

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... others (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, *32*.

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., ... others (2022). Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*. doi: https://doi.org/10.48550/arXiv.2202.11214

Rasp, S., Pritchard, M. S., & Gentine, P. (2018). Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences*, *115*(39), 9684–9689. doi: https://doi.org/10.1073/pnas.1810286115

Reynolds, R. W., Smith, T. M., Liu, C., Chelton, D. B., Casey, K. S., & Schlax, M. G. (2007). Daily high-resolution-blended analyses for sea surface temperature. *Journal of climate*, *20*(22), 5473–5496. doi: https://doi.org/10.1175/2007JCLI1824.1

Rodwell, M., & Palmer, T. (2007). Using numerical weather prediction to assess climate models. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, *133*(622), 129–146. doi: https://doi.org/10.1002/qj.23

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241). doi: https://doi.org/10.1007/978-3-319-24574-4_28

Rothrock, D., & Thorndike, A. (1984). Measuring the sea ice floe size distribution. *Journal of Geophysical Research: Oceans*, *89*(C4), 6477–6486. doi: https://doi.org/10.1029/JC089iC04p06477

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. doi: https://doi.org/10.48550/arXiv.1409.1556

Smith, G., Bélanger, J.-M., Roy, F., Pellerin, P., Ritchie, H., Onu, K., ... others (2018). Impact of coupling with an ice–ocean model on global medium-range NWP forecast skill. *Monthly Weather Review*, *146*(4), 1157–1180. doi: https://doi.org/10.1175/MWR-D-17-0157.1

Smith, M., Holland, M., & Light, B. (2022). Arctic sea ice sensitivity to lateral melting representation in a coupled climate model. *The Cryosphere*, *16*(2), 419–434. doi: https://doi.org/10.5194/tc-16-419-2022

Stevens, B., & Bony, S. (2013). What are climate models missing? *Science*, *340*(6136), 1053–1054. doi: https://doi.org/10.1126/science.1237554

Thorndike, A. S., Rothrock, D. A., Maykut, G. A., & Colony, R. (1975). The thickness distribution of sea ice. *Journal of Geophysical Research*, *80*(33), 4501–4513. doi: https://doi.org/10.1029/JC080i033p04501

Trémolet, Y. (2007). Incremental 4d-Var convergence study. *Tellus A: Dynamic Meteorology and Oceanography*, *59*(5), 706–718. doi: https://doi.org/10.1111/j.1600-0870.2007.00271.x

Tsamados, M., Feltham, D. L., & Wilchinsky, A. (2013). Impact of a new anisotropic rheology on simulations of arctic sea ice. *Journal of Geophysical Research: Oceans*, *118*(1), 91–107. doi: https://doi.org/10.1029/2012JC007990

Tsujino, H., Urakawa, S., Nakano, H., Small, R. J., Kim, W. M., Yeager, S. G., ... others (2018). JRA-55 based surface dataset for driving ocean–sea-ice models (JRA55-do). *Ocean Modelling*, *130*, 79–139. doi: https://doi.org/10.1016/j.ocemod.2018.07.002

Wang, C., Zhang, L., Lee, S.-K., Wu, L., & Mechoso, C. R. (2014). A global perspective on CMIP5 climate model biases. *Nature Climate Change*, *4*(3), 201–205. doi: https://doi.org/10.1038/nclimate2118

Wang, P., Yuval, J., & O'Gorman, P. A. (2022). Non-local parameteriza-

tion of atmospheric subgrid processes with neural networks. *Journal of Advances in Modeling Earth Systems*, *14*(10), e2022MS002984. doi: https://doi.org/10.1029/2022MS002984

Watt-Meyer, O., Brenowitz, N. D., Clark, S. K., Henn, B., Kwa, A., McGibbon, J., ... Bretherton, C. S. (2021). Correcting weather and climate models by machine learning nudged historical simulations. *Geophysical Research Letters*, *48*(15), e2021GL092555. doi: https://doi.org/10.1029/2021GL092555

Wergen, W. (1992). The effect of model errors in variational assimilation. *Tellus A*, *44*(4), 297–313. doi: https://doi.org/10.1034/j.1600-0870.1992.t01-3-00002.x

Yuval, J., & O'Gorman, P. A. (2020). Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions. *Nature communications*, *11*(1), 1–10. doi: https://doi.org/10.1038/s41467-020-17142-3

Zanna, L., & Bolton, T. (2020). Data-driven equation discovery of ocean mesoscale closures. *Geophysical Research Letters*, *47*(17), e2020GL088376. doi: https://doi.org/10.1029/2020GL088376

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision* (pp. 818–833). doi: https://doi.org/10.1007/978-3-319-10590-1_53

Zhang, C., Perezhogin, P., Gultekin, C., Adcroft, A., Fernandez-Granda, C., & L, Z. (2023). Implementation and evaluation of a machine learned mesoscale eddy parameterization into a numerical ocean circulation model. *arXiv preprint arXiv: 2303.00962*. doi: https://doi.org/10.48550/ARXIV.2303.00962

Zhang, Y., Bushuk, M., Winton, M., Hurlin, B., Delworth, T., Harrison, M., ... Yang, X. (2022). Subseasonal-to-seasonal Arctic sea ice forecast skill improvement from sea ice concentration assimilation. *Journal of Climate*, 1–48. doi: https://doi.org/10.1175/JCLI-D-21-0548.1

Zhang, Y., Bushuk, M., Winton, M., Hurlin, B., Yang, X., Delworth, T., & Jia, L. (2021). Assimilation of satellite-retrieved sea ice concentration and prospects for September predictions of Arctic sea ice. *Journal of Climate*, *34*(6), 2107–2126. doi: https://doi.org/10.1175/JCLI-D-20-0469.1

Zhu, Y., Zhang, R.-H., Moum, J. N., Wang, F., Li, X., & Li, D. (2022). Physics-informed deep learning parameterization of ocean vertical mixing improves climate simulations. *National Science Review*. doi: https://doi.org/10.1093/nsr/nwac044

Zupanski, M. (1993). Regional four-dimensional variational data assimilation in a quasi-operational forecasting environment. *Monthly weather review*, *121*(8), 2396–2408. doi: https://doi.org/10.1175/1520-0493(1993)121⟨2396:RFDVDA⟩2.0.CO;2

Supporting Information S1: Deep learning of systematic sea ice model errors from data assimilation increments

William Gregory[1], Mitchell Bushuk[2], Alistair Adcroft[1], Yongfei Zhang[1], Laure Zanna[3]

[1]Atmospheric and Oceanic Sciences Program, Princeton University, NJ, USA
[2]Geophysical Fluid Dynamics Laboratory, NOAA, Princeton, NJ, USA
[3]Courant Institute of Mathematical Sciences, New York University, New York, NY, USA

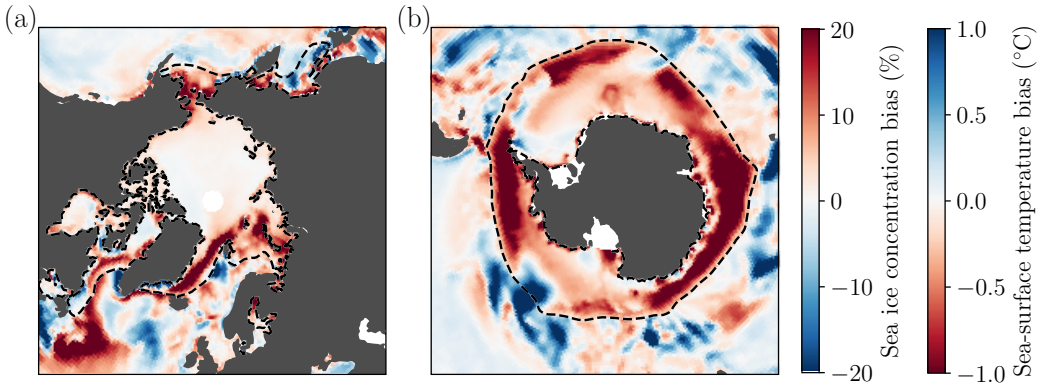

Figure S1. Seasonal climatologies of the SPEAR free-running model bias. Inside the climatological sea ice extent contour (black dashed line) are the aggregate sea ice concentration biases (model minus observations). Outside the contour are the sea-surface temperature biases (model minus observations). a) Arctic biases (December–February), b) Antarctic biases (September–November).

Corresponding author: Will Gregory, wg4031@princeton.edu

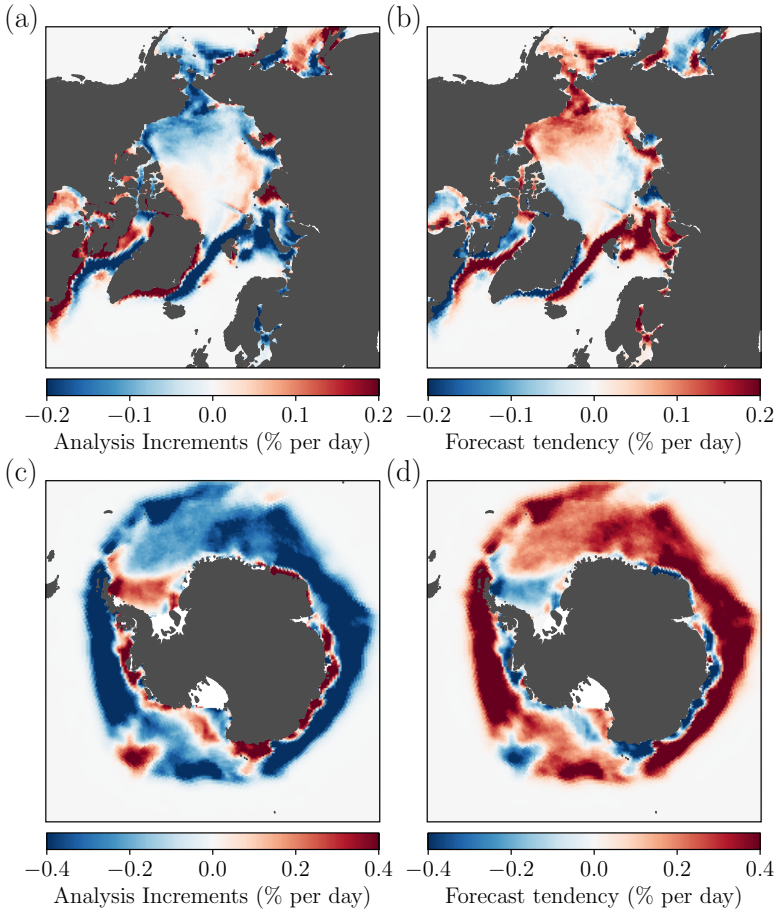

Figure S2. Comparisons of the climatological aggregate sea ice concentration analysis increments (a,c), and the aggregate sea ice concentration forecast tendencies (b,d). The spatial pattern correlation between panels a) and b) is -0.99. Similarly, the spatial pattern correlation between panels c) and d) is also -0.99.

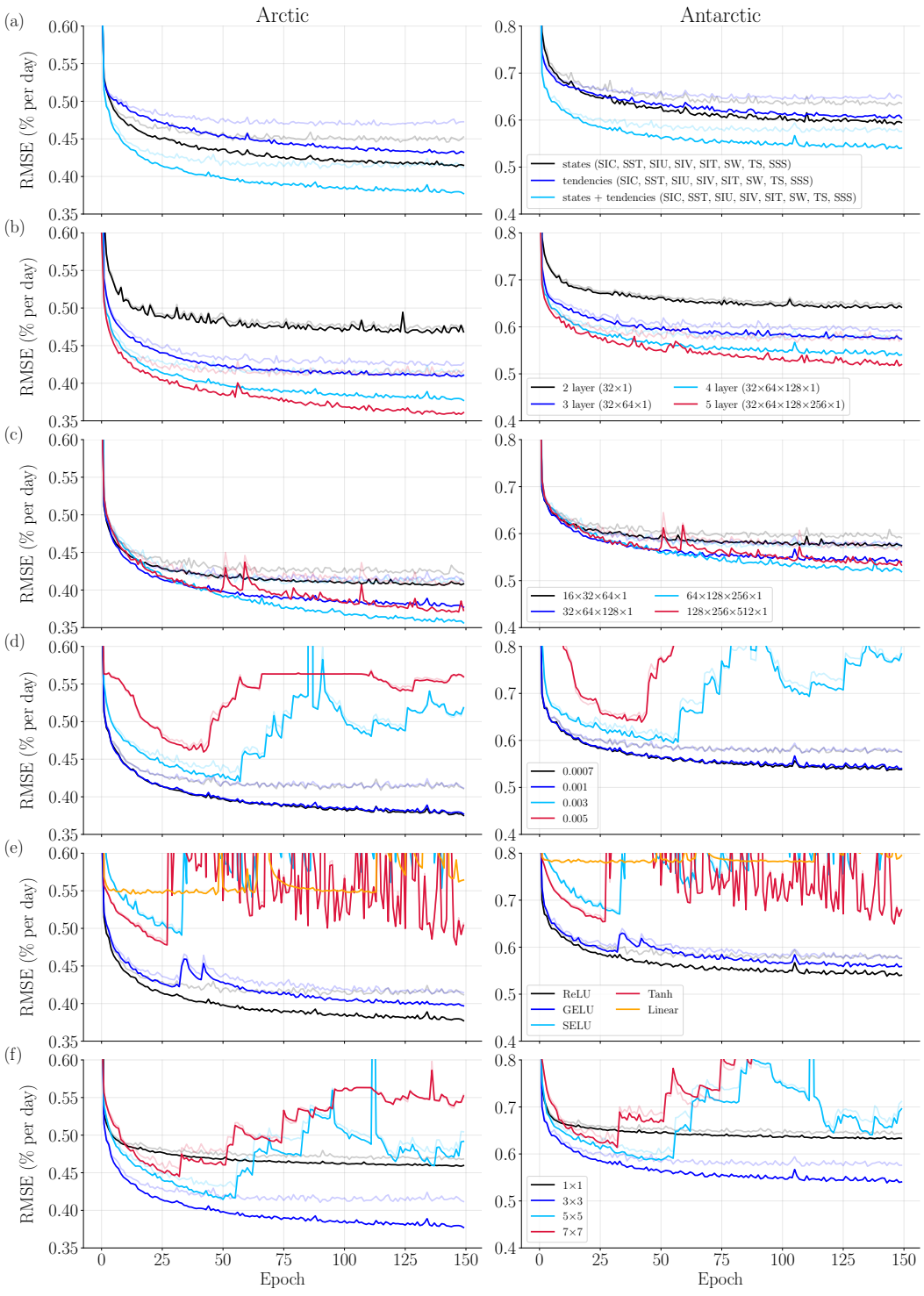

Figure S3. Learning curve examples for various CNN model selection tests. Each curve is the mean 5-fold cross-validation error on ΔSIC predictions (solid lines = error on training samples, transparent curves = error on validation samples). (a) Tests of the network sensitivity to the inputs (i.e., using just state variables, or just tendencies, or both). (b) Tests of the network depth (number of convolutional layers). (c) Tests of the network width (features per convolutional layer). (d) Tests of the optimizer learning rate. (e) Tests of the activation function used after each convolution operation. (f) Tests of the size of the convolution kernel used in each layer.

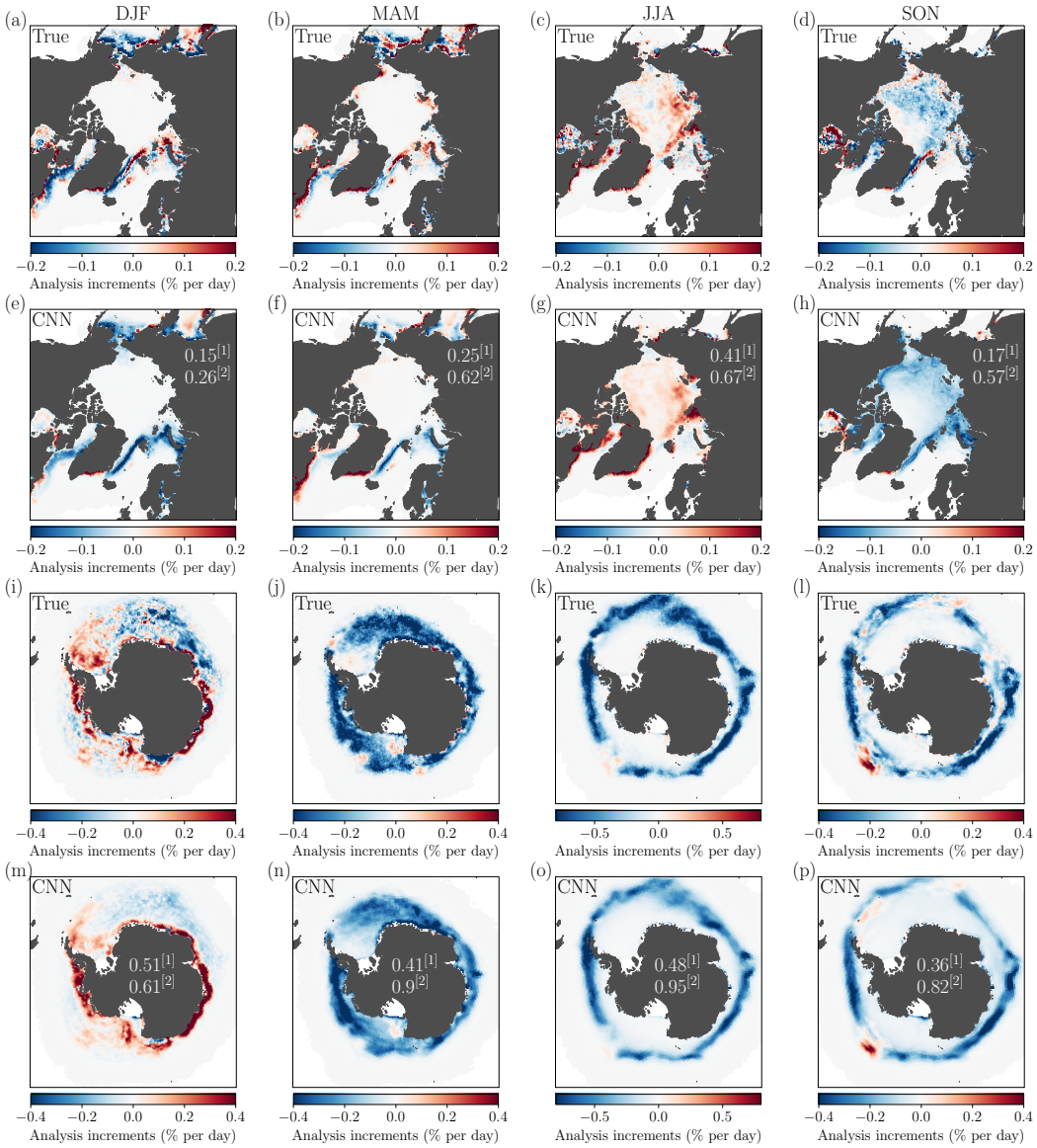

Figure S4. Seasonal climatologies of the (true) SPEAR category 1 sea ice concentration analysis increments and the equivalent CNN predictions, for both the Arctic (a)–(h) and Antarctic (i)–(p). Columns from left to right show DJF, MAM, JJA, SON climatologies, computed over the period 1982–2017. Values with the superscript [1] are the average of daily spatial pattern correlations between $\Delta\text{SICN}^{\text{True}}$ and $\Delta\text{SICN}^{\text{CNN}}$ in each respective season, while values with [2] are the spatial pattern correlations between the respective climatologies of the true and predicted increments.

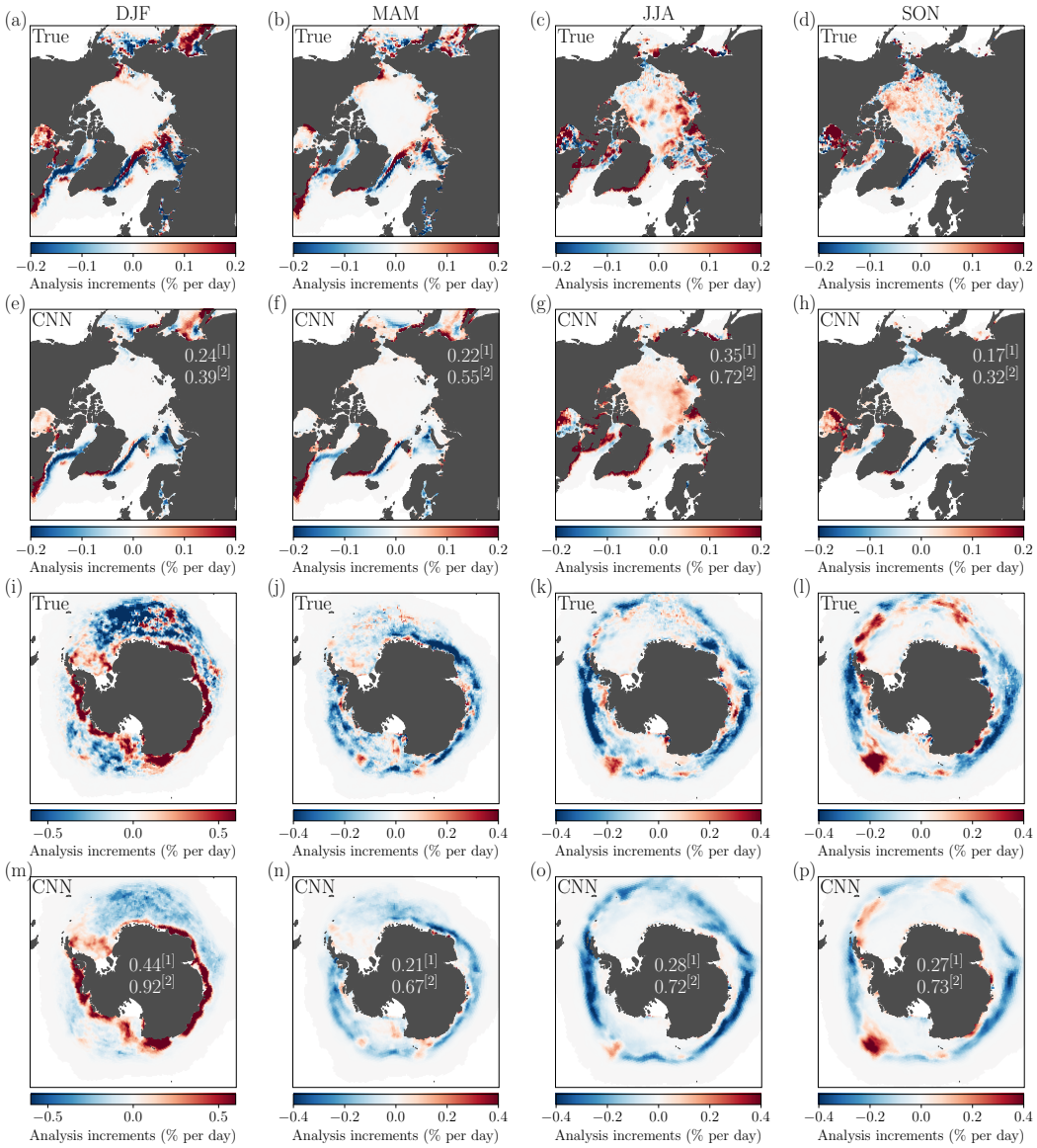

Figure S5. Seasonal climatologies of the (true) SPEAR category 2 sea ice concentration analysis increments and the equivalent CNN predictions, for both the Arctic (a)–(h) and Antarctic (i)–(p). Columns from left to right show DJF, MAM, JJA, SON climatologies, computed over the period 1982–2017. Values with the superscript [1] are the average of daily spatial pattern correlations between $\Delta\text{SICN}^{\text{True}}$ and $\Delta\text{SICN}^{\text{CNN}}$ in each respective season, while values with [2] are the spatial pattern correlations between the respective climatologies of the true and predicted increments.

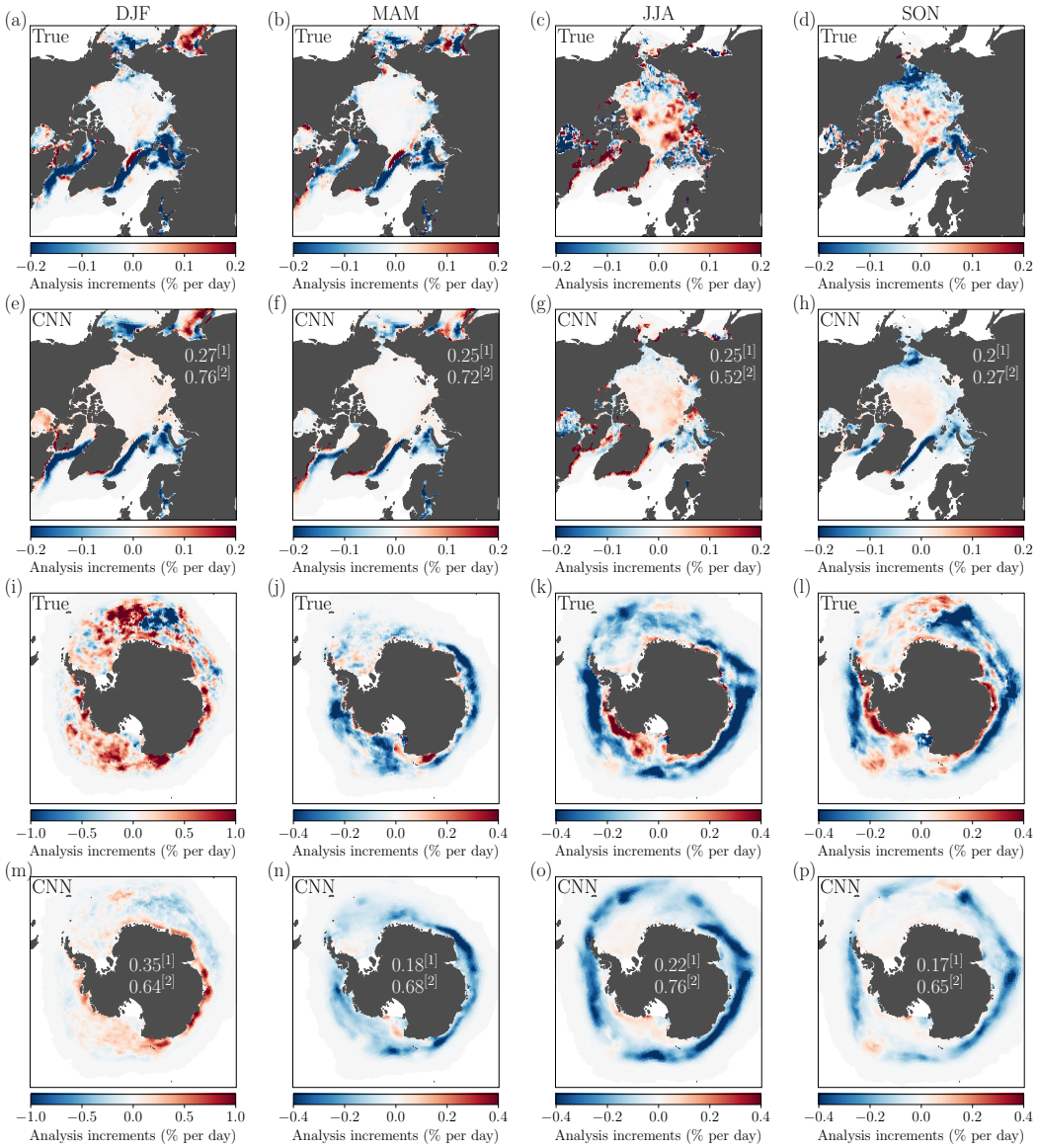

Figure S6. Seasonal climatologies of the (true) SPEAR category 3 sea ice concentration analysis increments and the equivalent CNN predictions, for both the Arctic (a)–(h) and Antarctic (i)–(p). Columns from left to right show DJF, MAM, JJA, SON climatologies, computed over the period 1982–2017. Values with the superscript [1] are the average of daily spatial pattern correlations between $\Delta\text{SICN}^{\text{True}}$ and $\Delta\text{SICN}^{\text{CNN}}$ in each respective season, while values with [2] are the spatial pattern correlations between the respective climatologies of the true and predicted increments.

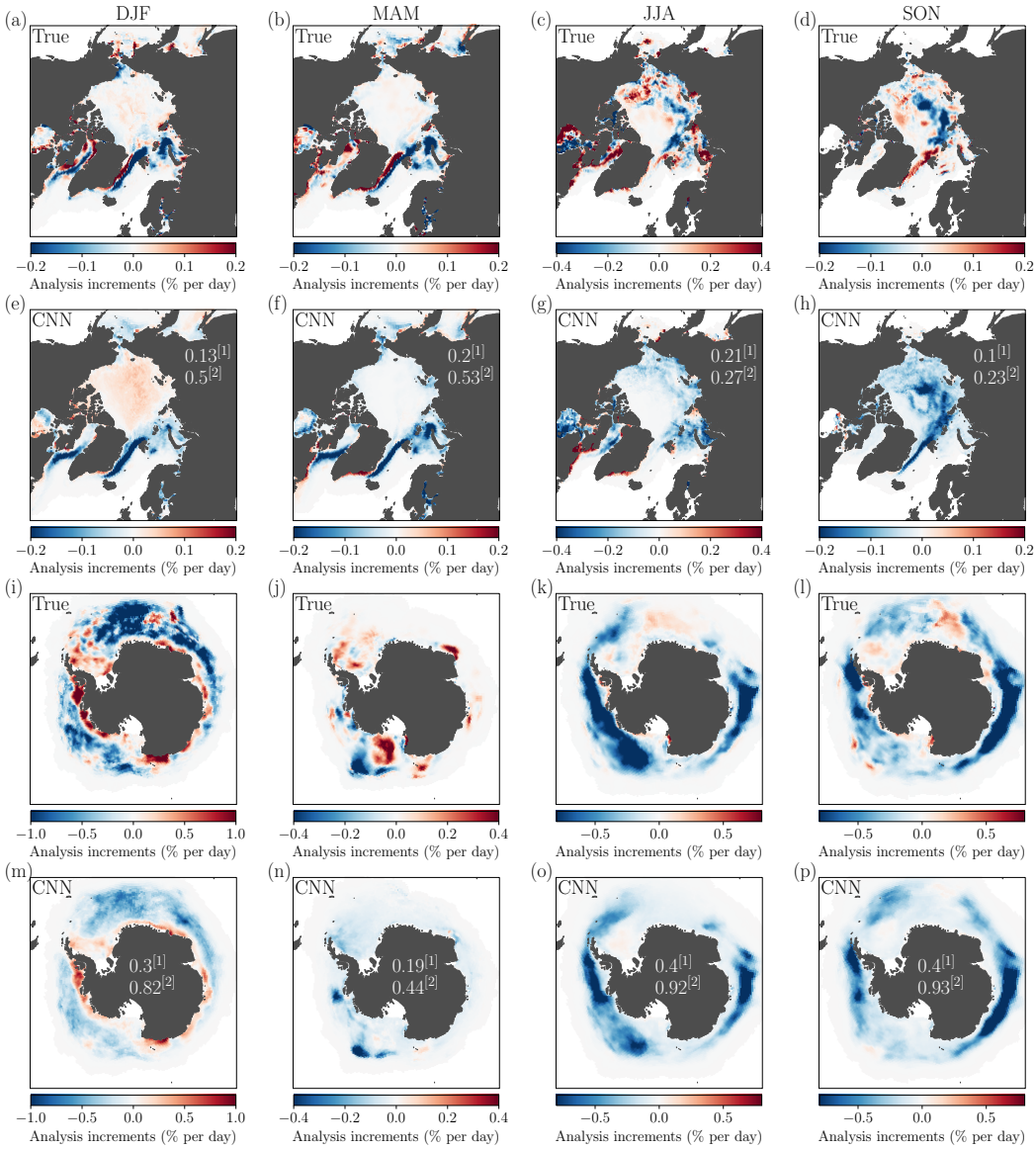

Figure S7. Seasonal climatologies of the (true) SPEAR category 4 sea ice concentration analysis increments and the equivalent CNN predictions, for both the Arctic (a)–(h) and Antarctic (i)–(p). Columns from left to right show DJF, MAM, JJA, SON climatologies, computed over the period 1982–2017. Values with the superscript [1] are the average of daily spatial pattern correlations between $\Delta\text{SICN}^{\text{True}}$ and $\Delta\text{SICN}^{\text{CNN}}$ in each respective season, while values with [2] are the spatial pattern correlations between the respective climatologies of the true and predicted increments.

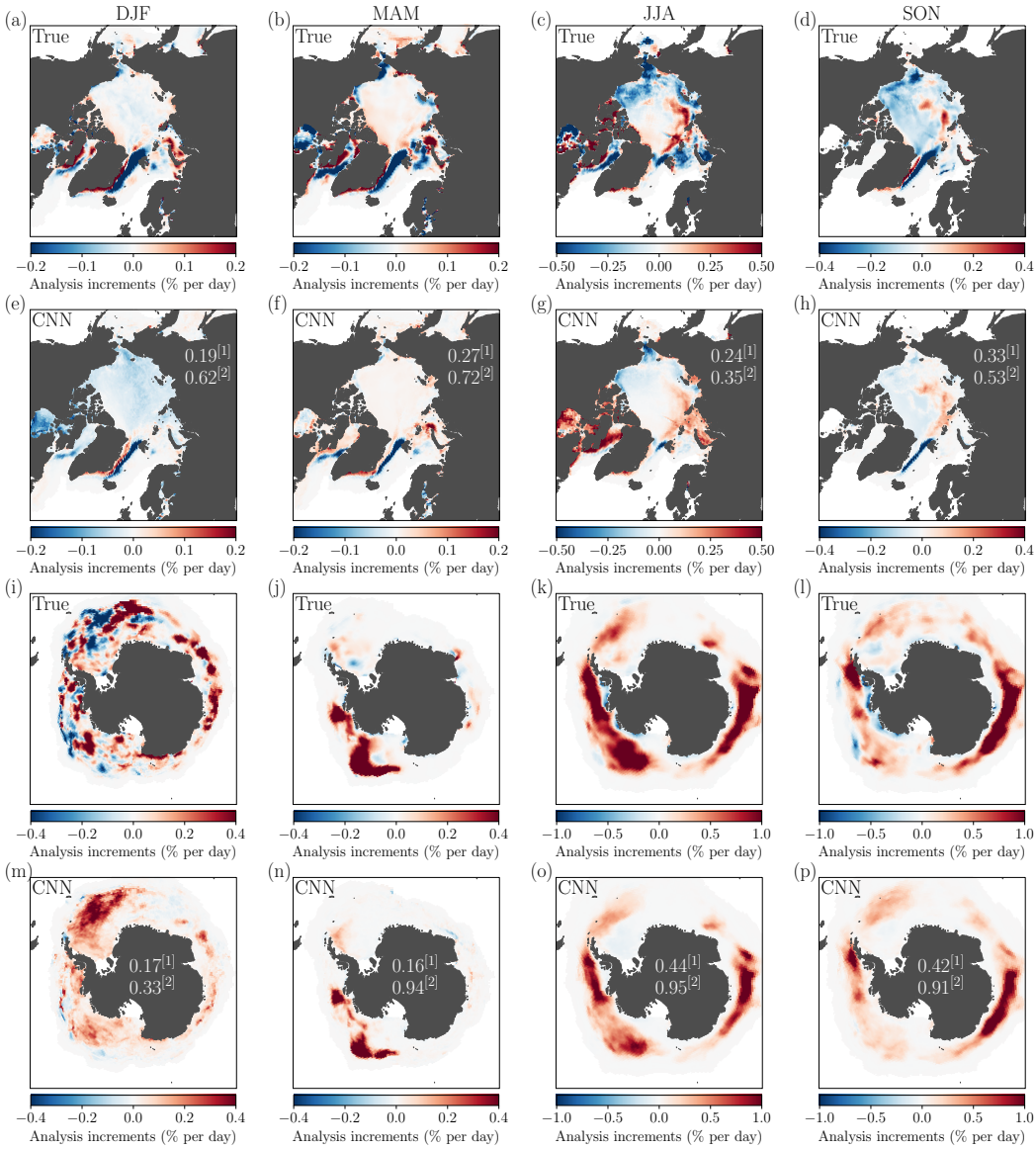

Figure S8. Seasonal climatologies of the (true) SPEAR category 5 sea ice concentration analysis increments and the equivalent CNN predictions, for both the Arctic (a)–(h) and Antarctic (i)–(p). Columns from left to right show DJF, MAM, JJA, SON climatologies, computed over the period 1982–2017. Values with the superscript [1] are the average of daily spatial pattern correlations between $\Delta\text{SICN}^{\text{True}}$ and $\Delta\text{SICN}^{\text{CNN}}$ in each respective season, while values with [2] are the spatial pattern correlations between the respective climatologies of the true and predicted increments.

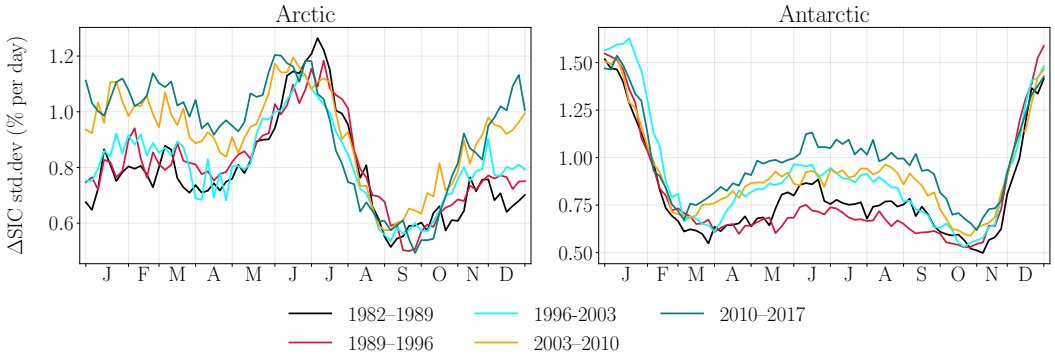

Figure S9. The standard deviation of the true aggregate sea ice concentration analysis increments ($\Delta\text{SIC}^{\text{True}}$), computed over each of the 5 cross-validation periods used for validating the CNN predictions. Shown as daily climatologies.

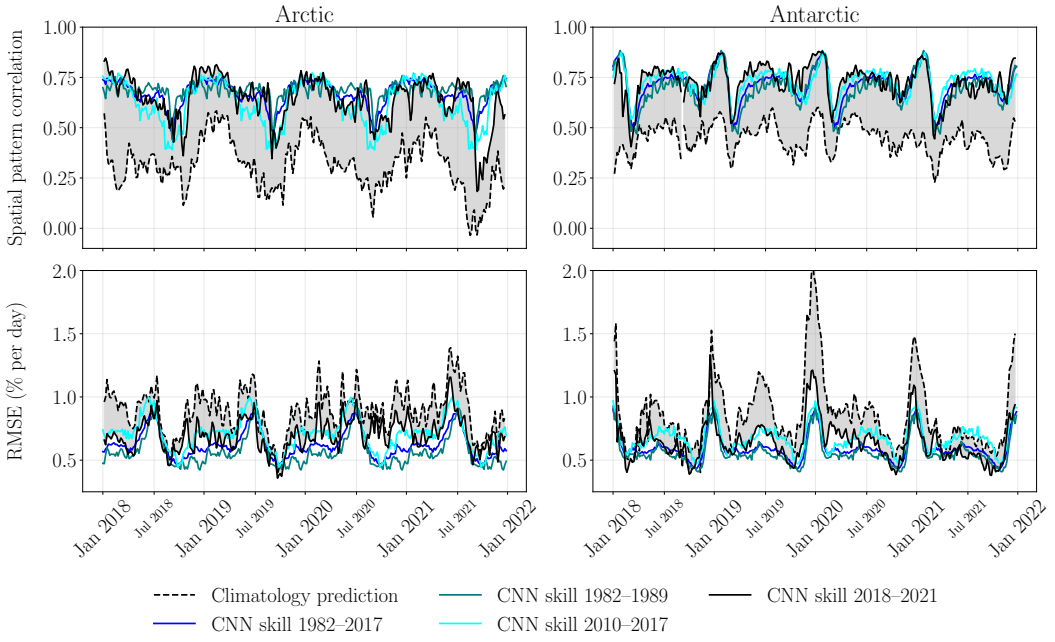

Figure S10. As in Figure 8 from the main article, except now highlighting the effect of non-stationarity within the increments by also including the climatological prediction skill for cross-validation chunks corresponding to the beginning of the time series record (1982–1989), as well as the end of the time series record (2010–2017).