Geophysical Research Letters

RESEARCH LETTER
10.1029/2023GL106776

Machine Learning for Online Sea Ice Bias Correction Within Global Ice-Ocean Simulations

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

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

Key Points:

- We use a convolutional neural network (CNN) to perform online sea ice bias correction within global ice-ocean simulations
- The CNN systematically reduces the free-running model bias in both the Arctic and Antarctic
- The online performance can be improved by combining CNN and data assimilation corrections in order to iteratively augment the training data

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

W. Gregory,
wg4031@princeton.edu

Citation:

Gregory, W., Bushuk, M., Zhang, Y., Adcroft, A., & Zanna, L. (2024). Machine learning for online sea ice bias correction within global ice-ocean simulations. *Geophysical Research Letters*, *51*, e2023GL106776. https://doi.org/10.1029/2023GL106776

Received 10 OCT 2023
Accepted 22 JAN 2024

Abstract In this study, we perform online sea ice bias correction within a Geophysical Fluid Dynamics Laboratory global ice-ocean model. For this, we use a convolutional neural network (CNN) which was developed in a previous study (Gregory et al., 2023, https://doi.org/10.1029/2023ms003757) for the purpose of predicting sea ice concentration (SIC) data assimilation (DA) increments. An initial implementation of the CNN shows systematic improvements in SIC biases relative to the free-running model, however large summertime errors remain. We show that these residual errors can be significantly improved with a novel sea ice data augmentation approach. This approach applies sequential CNN and DA corrections to a new simulation over the training period, which then provides a new training data set to refine the weights of the initial network. We propose that this machine-learned correction scheme could be utilized for generating improved initial conditions, and also for real-time sea ice bias correction within seasonal-to-subseasonal sea ice forecasts.

Plain Language Summary Climate models contain errors which often lead to predictions which are consistently out of agreement with what we observe in reality. In some cases we know the origin of these errors, for example, predicting too much sea ice as a result of consistently cool ocean temperatures. In reality, however, there are typically numerous model errors interacting across the atmosphere, ocean and sea ice, and to manually parse through large volumes of climate model data in an attempt to isolate these errors in time and space is highly impractical. Machine learning on the other hand is a framework which is well-suited to this task. In this work we take a machine learning model which, at any given moment, ingests information about a climate model's atmosphere, ocean and sea ice conditions, and predicts how much error there is in the climate model's representation of sea ice, without seeing any actual sea ice observations. We use this to adjust the sea ice conditions in one particular climate model as it is running forward in time making predictions, and we find that this significantly reduces the model's sea ice errors globally.

1. Introduction

Machine learning (ML) algorithms are beginning to cement their position as viable subgrid-scale climate model parameterizations, through their ability to isolate complex non-linear relationships within large volumes of high dimensional data (Brenowitz & Bretherton, 2018; Gentine et al., 2018; O'Gorman & Dwyer, 2018; Sane et al., 2023; Yuval & O'Gorman, 2020). Typically this is achieved by training an ML model to learn a functional mapping which characterizes the impact of subgrid processes on resolved scales, by training on high resolution simulations or observational data. Significant effort is currently being afforded to the development of these ML parameterizations in the context of for example, ocean turbulence, with early results (Frezat et al., 2022; Kurz et al., 2023; Ross et al., 2023; Zanna & Bolton, 2020; C. Zhang et al., 2023) highlighting their potential to improve important climate statistics, such as eddy kinetic energy at large scales, over their traditional physics-based counterparts. Meanwhile, only recently have studies begun to investigate ML-based subgrid sea ice parameterizations. For example, Finn et al. (2023) successfully demonstrated an ML-based sea ice parameterization which learns short timescale errors associated with subgrid sea ice dynamics in a low resolution sea ice model. Furthermore, Driscoll et al. (2023) showed success in emulating a sea ice melt pond parameterization using neural networks.

Alternatively, combining data assimilation (DA) and ML has shown to be a promising framework for learning either subgrid parameterizations or systematic model errors across various domains (Arcucci et al., 2021; Bonavita & Laloyaux, 2020; Brajard et al., 2021; Chen et al., 2022; Farchi et al., 2021, 2023; He et al., 2023;

Laloyaux et al., 2022; Mojgani et al., 2022). Readers are also referred to Bocquet et al. (2020) and Brajard et al. (2020) for seminal works in this field. In a recent sea ice study by Gregory et al. (2023), hereafter G23, the authors presented a DA-based ML framework in which convolutional neural networks (CNNs) were used to predict state-dependent sea ice errors within an ice-ocean configuration of the Geophysical Fluid Dynamics Laboratory (GFDL) Seamless system for Prediction and EArth System Research (SPEAR) model, as a way to highlight the feasibility of a data-driven sea ice model parameterization within SPEAR. They approached this by first showing that the climatological sea ice concentration analysis increments from an ice-ocean DA experiment map closely onto the systematic bias patterns of the equivalent free-running model. This suggested that an ML model which is able to predict the analysis increments could, in principle, reduce sea ice biases as an online model parameterization or bias correction tool. Their subsequent CNN architecture then used information from local model state variables and their tendencies, to make predictions of the corresponding sea ice concentration analysis increment at any grid cell location. These offline predictions were shown to generalize well to both the Arctic and Antarctic domains, and across all seasons. However, offline performance does not always directly translate to online simulations, which can sometimes exhibit instabilities as well as climate drift after implementation (Brenowitz et al., 2020; Ott et al., 2020; Rasp et al., 2018). In such cases, the ML model may require an additional online training step in order to sample a larger model state space to which it was initially trained (Rasp, 2020).

In this present work, we advance the field of ML-based parameterization in sea ice modeling by investigating the online performance of the G23 DA-based ML model when used as a tool to correct short-term sea ice error growth. We implement the correction scheme here within a coupled ice-ocean configuration of SPEAR, as the G23 CNN was originally trained on data from an ice-ocean DA system, which therefore allows us to make direct comparisons of model biases and increments produced from both the CNN and DA simulations. If the CNN is able to reduce sea ice biases relative to the free-running model, then this will provide a solid foundation for future work into assessing the generalization to fully coupled systems, and ultimately a physics-based sea ice model parameterization.

2. Data and Methods

2.1. SPEAR Ice-Ocean Model

SPEAR is a fully coupled ice-ocean-atmosphere-land model (Delworth et al., 2020), which shares the same components as the GFDL CM4 model (Held et al., 2019), however with parameterizations and resolutions geared toward seasonal-to-decadal prediction. The ocean and sea ice components are configured at a 1° horizontal resolution and correspond to the Modular Ocean Model v6 (MOM6) and the Sea Ice Simulator v2 (SIS2), respectively (Adcroft et al., 2019). In this work, we consider an ice-ocean configuration of SPEAR, in which MOM6 and SIS2 are forced by atmospheric conditions from the Japanese 55-year Reanalysis for driving ocean-sea-ice models (JRA55-do; Tsujino et al. (2018)). Details of the ice-ocean experiments are provided in Section 2.3.

2.2. Machine Learning Model

The CNN model from G23 was trained to predict sea ice concentration (SIC) increments from a SPEAR ice-ocean DA experiment (Y. Zhang et al. (2021); hereafter Z21). The Z21 DA experiment spanned 1 January 1982–1 January 2018, where satellite observations of SIC from the National Snow and Ice Data Center (NSIDC; Cavalieri et al. (1996)) NASA Team algorithm were assimilated into SIS2 every 5 days using the Ensemble Adjustment Kalman Filter (EAKF) approach (Anderson, 2001), and sea-surface temperatures were nudged toward observations from version 2 of the Optimum Interpolation Sea-Surface Temperature (OISSTv2) data set (Banzon et al., 2016; Reynolds et al., 2007) at the model timestep. It should be noted that SIS2 has a 5-category ice thickness distribution (Bitz et al., 2001), with lower thickness bounds of 0.0, 0.1, 0.3, 0.7, and 1.1 m. The (observable) aggregate SIC field is therefore computed in the model as the sum of the sea ice concentration in each category (SICN), hence $\mathrm{SIC} = \sum_{k=1}^{5} \mathrm{SICN}_k$. Similarly, we compute the aggregate SIC increment (ΔSIC) as the sum of the analysis increments in each category (ΔSICN). The G23 CNN then uses 5-day mean inputs of state variables and tendencies corresponding to: SIC, sea-surface temperature (SST), zonal and meridional components of ice velocities, sea ice thickness, net shortwave radiation, ice-surface skin temperature, sea-surface salinity, and a land-sea mask, in order to predict ΔSIC. This prediction of ΔSIC is then passed to a second CNN, along with

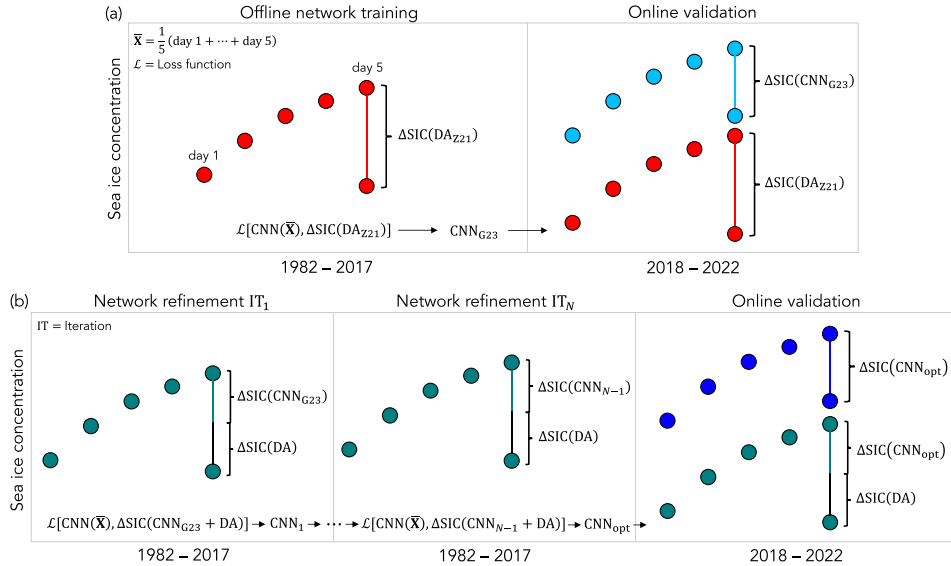

Figure 1. Schematic of bias correction schemes, shown in each panel for one 5-day assimilation/correction cycle. The dots represent the daily model state integrating forward in time. The vertical lines are then the corrections from either the CNN or DA. (a) Out-of-the-box G23 CNN training and implementation showing the Z21 DA simulation in red and the CNN simulation in light blue. (b) Optimized G23 CNN training and implementation showing simulations with combined CNN and DA corrections in teal, and the optimized CNN simulation in dark blue.

SICN, to predict the category concentration increments ΔSICN. For convenience we refer to these two CNNs as a single network hereafter.

The implementation of the CNN into SIS2 here is performed in an analogous manner to DA. Specifically, we run an ensemble forecast of the model for 5 days (e.g., from 00:00 hr UTC on 1 January to 00:00 UTC on 6 January), where we then generate the corresponding ΔSICN predictions for each ensemble member, add the predicted ΔSICN fields to the instantaneous SICN state (i.e., the state at 00:00 UTC on 6 January), and restart the model for the next 5-day forecast (schematics of this 5-day forecast plus correction process are shown in Figure 1, although Section 2.3 describes this figure in more detail). It is important to note that we also apply a post-processing after each correction. For this, we follow the Z21 procedure for updating sea ice variables during DA (see Text S1 in Supporting Information S1).

2.3. Ice-Ocean Experiments

We compare four ice-ocean simulations in this study, where each extends for a 5-year period between 1 January 2018 and 1 January 2023. The initial ice and ocean conditions for all simulations are based on those from the Z21 DA experiment, which ended 1 January 2018. The atmospheric forcing is provided by JRA55-do reanalysis version 1.5, SSTs are nudged toward OISSTv2 observations using a piston velocity of 4 m per day, and SSS is nudged to a seasonal climatology with a piston velocity of 1/6 m per day. Note that "piston velocity" here follows previous definitions (Griffies et al., 2009), and refers to a scaling of the associated heat or salt flux due to restoring the modeled ocean temperature or salinity to observations, respectively. Larger piston velocities yield a larger heat or salt flux for a given temperature or salinity difference, respectively. In any case, the experiments are given as follows:

1. The free-running model in ice-ocean mode (FREE).
2. An extension of the ice-ocean DA experiment (DA_{Z21}), which serves as the benchmark for this study.
3. An "out-of-the-box" implementation of the G23 network (CNN_{G23}), where the network has been trained offline using all available data from the original DA experiment. This procedure is highlighted in Figure 1a, where, during training, the loss function $\mathcal{L}$ minimizes the error between the network predictions, $\text{CNN}(\bar{\mathbf{X}})$, and the increment from DA, $\Delta\text{SIC}(\text{DA}_{\text{Z21}})$. Here $\bar{\mathbf{X}}$ represents the 5-day mean state variables and tendencies described in Section 2.2. After training, this produces the network CNN_{G23}, which is then implemented over

the 2018–2022 period. The reader is referred to G23 for more details of the architecture and hyperparameters related to the network training process.

4. An "optimized" version of the G23 network (CNN_{opt}) where the weights of the G23 network are refined to improve online performance. For this we use both DA and CNNs to iteratively augment the training data, and subsequently refine the network weights after each augmentation iteration. For example, in the first iteration we run a new ice-ocean simulation between 1982 and 2017, and in which we apply a two-step CNN + DA correction every 5 days; first using CNN_{G23}, and then using DA (see Figure 1b). We then use this 36-year simulation as a new training data set with which to update the weights of CNN_{G23}, where, during training, the loss function now minimizes the error between the network predictions, $CNN(\tilde{\mathbf{X}})$, and the total model error, $\Delta SIC(CNN_{G23} + DA)$. This procedure is performed for a total of $N = 3$ iterations. Note that we use the term "augmentation" here as each iteration corresponds to an update of network weights from the previous iteration, rather than updating from for example, randomly initialized weights. Therefore, CNN_{opt} has been exposed to 3X more data samples than CNN_{G23}. The network refinement after each augmentation iteration is performed in an identical way to the offline learning procedure outlined in G23, except now we only update the weights for five epochs after each iteration. After the three iterations, the final optimized network CNN_{opt} is then implemented over the 2018–2022 period without DA to assess performance (blue dots in Figure 1b).

Note that the "Online validation" panels in Figure 1 highlight the simulations with CNN implementations relative to a simulation which applies the respective "perfect" correction (i.e., the correction which either CNN would produce if it had 100% prediction accuracy). These simply correspond to the extended DA experiment for CNN_{G23} in Figure 1a, and the two-step CNN correction plus DA for CNN_{opt} in Figure 1b. A comparison of these corrections (increments) is made in Section 3.2.2 in order to establish how the different implementation configurations manifest within the increments. As a final point to note here, all results presented in this work are based on ensemble mean fields, and all simulations are run with a "no leap" calendar, which excludes leap-year days.

3. Results

3.1. Model Bias

Figure 2 shows model biases for each of the ice-ocean experiments outlined in Section 2.3. Initially considering the annual-mean spatial bias patterns of the free-running model, we can see that this simulation is overall positively biased in both hemispheres, with largest Arctic biases occurring in the east Atlantic sector (Greenland, Barents, and Kara seas), and largest Antarctic biases in the Bellingshausen, Amundsen, and Indian Ocean sectors. The daily SIC root-mean-squared error (RMSE) curves then highlight the seasonal variation of the model bias, with largest RMSE values in FREE (black curves) occurring across June–August in the Arctic (22.7%), and December–February in the Antarctic (28.5%). The average RMSE over the entire simulation period corresponds to 17.8% and 20.0% in the Arctic and Antarctic, respectively, with larger errors in summer and smaller errors in winter. As expected, the DA experiment (DA_{Z21}; red curves) visibly reduces the bias across all seasons, with average Arctic and Antarctic RMSE reductions relative to FREE of 4.2% and 5.5%, respectively.

Turning to the two CNN correction schemes, the out-of-the-box implementation (CNN_{G23}; light blue curves) shows systematic improvements relative to FREE, with average RMSE reductions of 1.9% and 3.9% in the Arctic and Antarctic, respectively. The modest improvements in the Arctic make it difficult to identify qualitative differences in the climatology spatial bias plots, however some improvements can be seen in the east Atlantic sector. This is also highlighted in the regional Arctic sea ice extent (SIE) time series (Figure S1 in Supporting Information S1), where regions are defined according to Meier and Stewart (2023). On the other hand, CNN_{G23} shows visible improvements across much of the Antarctic domain, with large bias reductions in the Amundsen Sea and Pacific Ocean. The regional Antarctic SIE time series (Figure S2 in Supporting Information S1) also more closely track the DA experiment throughout the majority of the simulation period, particularly in the Antarctic growth season. In the melt season however, the simulation shows a tendency to drift back toward to the free-running model state. Comparing these results to the optimized CNN implementation (CNN_{opt}; dark blue curves), we see marked skill improvements. The average RMSE reductions compared to FREE are 3.9% and 5.9% in the Arctic and Antarctic, respectively, and Figure 2 shows that sizable improvements have been made in the

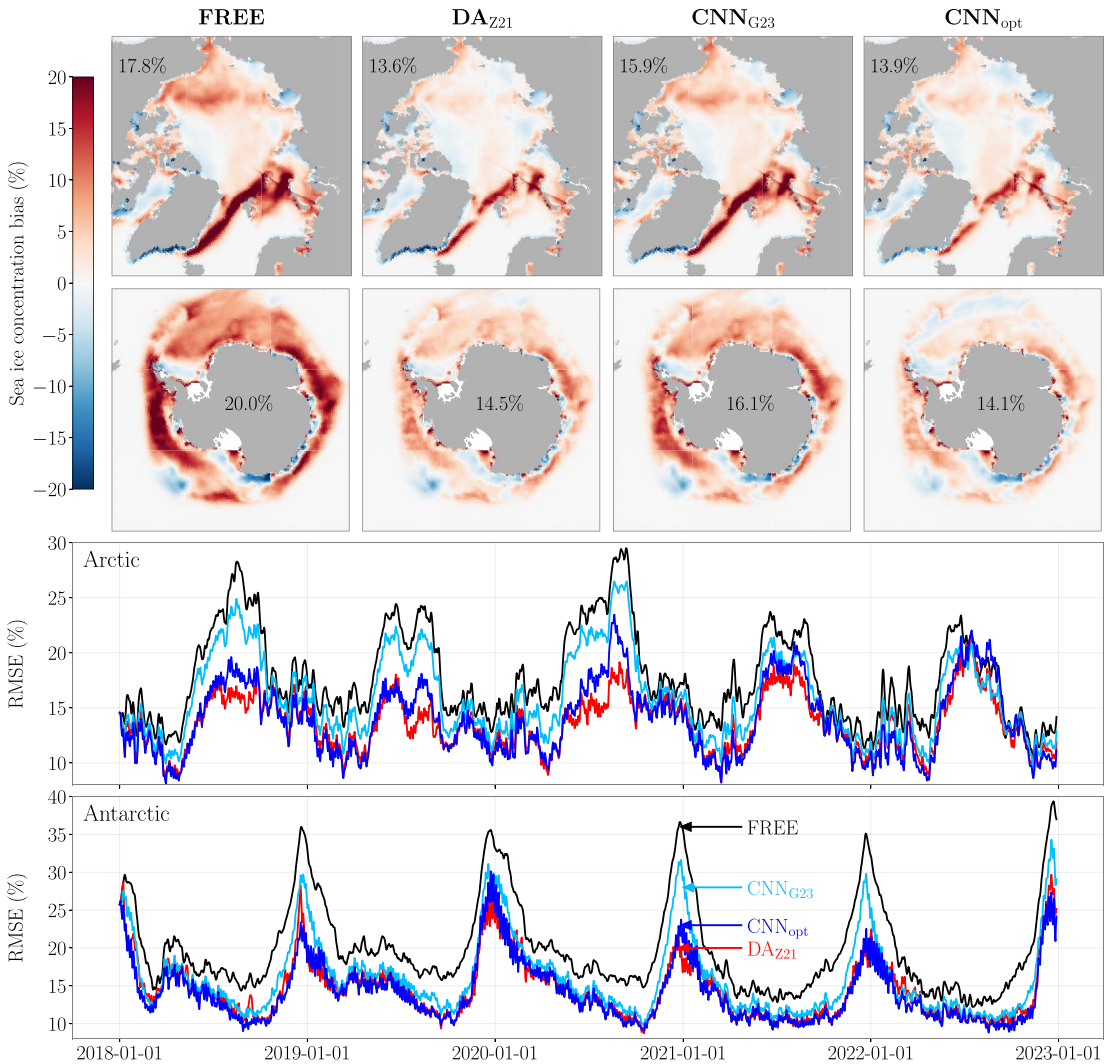

Figure 2. Comparison of model biases from ice-ocean simulations over the 2018–2022 period. The first and second rows show the mean SIC biases (model minus observations), relative to the NSIDC NASA Team observations (DiGirolamo et al., 2022). The average RMSE of SIC is reported in each panel (RMSE computed each day over sea ice covered grid cells only, and then averaged over all days). The bottom two time series plots show pan-Arctic and pan-Antarctic RMSE of SIC for each simulation.

summer months in both hemispheres. Furthermore, both pan-Antarctic and regional SIE (Figure S2 in Supporting Information S1) are also considerably improved in the melt season compared to CNN$_{G23}$, and often show reduced biases relative to the DA experiment. It is worth noting however that many of the regional Antarctic SIE time series for CNN$_{opt}$ show visible imprints of model shock (i.e., large fluctuations in extent, occurring every 5 days). This can occur in DA when there is significant drift between each assimilation cycle, and the fact that we see this here may suggest that there is rapid error growth occurring over the space of 5 days in the Antarctic. We discuss this further in Section 3.2.2. In any case, the fact that the CNN$_{opt}$ experiment, which does not assimilate any observations, has similar errors to DA$_{Z21}$ suggests that the DA run is primarily correcting systematic model error and that CNN$_{opt}$ is successfully capturing these errors.

3.2. Understanding Online Improvements

Between the two CNN models, it is clear that CNN$_{opt}$ is the most desirable scheme for reducing the free-running model bias. Furthermore, it is also clear that, relative to CNN$_{G23}$, the largest gains from CNN$_{opt}$ come in the summer months. In this section we take a closer look at the performance of each CNN correction scheme in order

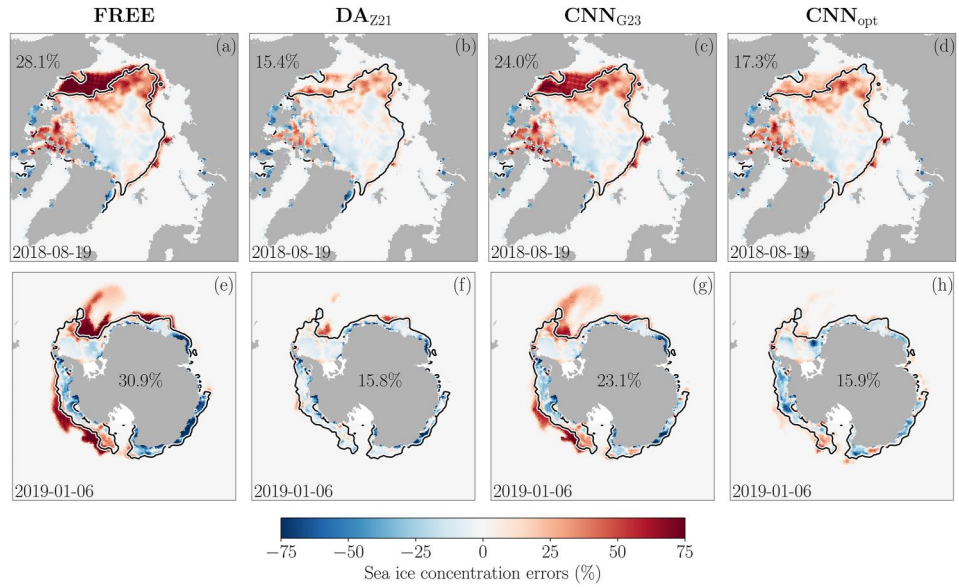

Figure 3. Snapshots of summertime model errors (model minus observations) for the FREE, DA and CNN ice-ocean simulations. Errors are computed relative to the NSIDC NASA Team observations (DiGirolamo et al., 2022). RMSE values are reported in each panel. The black contours mark the observed sea ice edge position (15% SIC).

to discern how these improvements manifest in both the spatial error patterns of each simulation, and also in the SIC increments produced from each scheme.

3.2.1. Snapshots

Figure 3 shows example snapshots of summertime model errors in each hemisphere (see also Movie S1 in Supporting Information S1 for snapshots over the 5-year simulation period). In both hemispheres we can see that FREE (Figures 3a and 3e) contains large positive errors related to over-estimation of the sea ice edge (indicated by the positive SIC biases equator-ward of the observed ice edge contour). The errors pole-ward of the ice edge contour indicate local SIC errors. While the DA simulation (Figures 3b and 3f) retains some of these local SIC errors, a significant fraction of the ice edge errors are reduced; almost halving the RMSE in the Antarctic relative to FREE. CNN$_{G23}$ (Figures 3c and 3g) shows some improvements relative to FREE (4.1% and 7.8% RMSE improvement in the Arctic and Antarctic, respectively), however there are still considerable ice edge and local SIC errors throughout both the Pacific sector in the Arctic, and the Atlantic and Pacific sectors in the Antarctic.

For CNN$_{opt}$ (Figures 3d and 3h) there are clear RMSE improvements relative to both FREE and CNN$_{G23}$ in both hemispheres, with remarkable improvements in the Antarctic. It could be argued however that, in the Arctic, the simulated ice edge position is not much improved in this example. A useful metric to confirm this is the integrated ice edge error (IIEE; Goessling et al. (2016)), which computes the total area for which the ice edge is both over- and under-predicted, relative to satellite observations. For panels (a–d) in Figure 3, the IIEEs are given as 1.44, 0.90, 1.32, 1.05 million km^2, respectively, which shows that the sea ice edge from each correction scheme is in better agreement with the observations than FREE, and that CNN$_{opt}$ does indeed improve over CNN$_{G23}$ in this regard. Similarly in the Antarctic panels (e–h), the IIEEs are given as 3.98, 1.38, 3.11, 1.30 million km^2, respectively. Here we can see that CNN$_{opt}$ even shows improved ice edge errors over the DA simulation (see Figure S3 in Supporting Information S1 for IIEE metrics computed over the entire 2018–2022 period). We can also take the assessment of ice edge errors further by disaggregating the SIC RMSE metric into grid points that lie pole-ward and equator-ward of the observed ice edge contour on any given day, in order to assess where the largest improvements from each correction scheme are manifesting (i.e., whether improvements are primarily in the ice edge location or SIC within the ice pack). From this decomposition (Figures S4 and S5 in Supporting Information S1) we find that, relative to FREE, in both hemispheres the largest RMSE reductions from each correction scheme come from improvements in the ice edge. Furthermore, we find that CNN$_{opt}$ is considerably reducing the summer ice edge errors relative to CNN$_{G23}$.

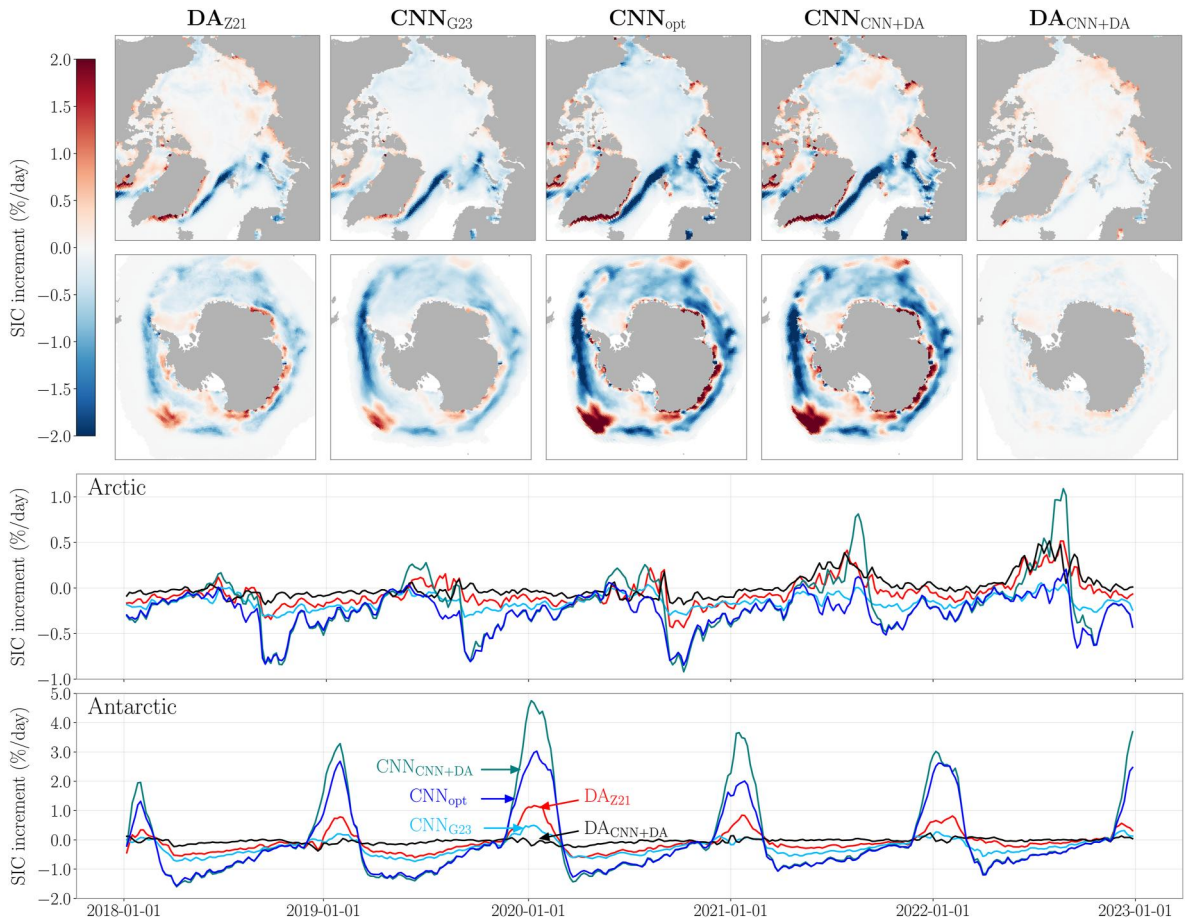

Figure 4. Comparison of SIC increments produced from either DA or CNNs during online simulations. The first two rows show spatial climatologies over 2018–2022. The bottom two panels are then the equivalent time series, computed as mean fields over Arctic and Antarctic domains.

3.2.2. Analysis Increments

Figure 4 shows the mean SIC increments for all DA and CNN correction schemes. The increments here correspond to those which were originally outlined in the "Online validation" panels of Figure 1. Namely, the extended DA experiment (DA_{Z21}), the out-of-the-box CNN implementation (CNN_{G23}), the optimized version of the G23 network (CNN_{opt}), and each of the corrections from a two-step CNN + DA process, in which CNN_{opt} provides an initial correction to the model state, followed immediately by a second correction from DA. We refer to these individual corrections in the CNN + DA run as CNN_{CNN+DA} and DA_{CNN+DA}, respectively. It should be noted that the simulations corresponding to CNN_{opt} and CNN_{CNN+DA} use identical networks over the 2018–2022 period, however, their inputs, and therefore their corrections, differ. Also, we note that there is no DA involved in the 2018–2022 simulation corresponding to CNN_{opt}.

The mean increments from both CNN_{G23} and CNN_{opt} in Figure 4 show largely similar spatial patterns in both hemispheres, with CNN_{opt} displaying overall larger magnitudes. In the Arctic, while both sets of CNN increments show isolated regions of positive values along the Eurasian coast, they do not reflect the larger area of mean positive increments seen in DA_{Z21} across the East Siberian, Laptev and Kara seas (from the DA_{Z21} Arctic time series panel in Figure 4 we can see that these positive increments originate in summer). On the other hand, the increments from CNN_{CNN+DA} do indeed show mean positive summer values in these regions. This suggests that the combination of input variables to the network which are needed to generate these positive predictions in the Arctic, is only being sampled after the additional DA_{CNN+DA} step. Nonetheless, in Section 3.2.1 we have seen that CNN_{opt} yields significant improvements in Arctic summer SIC errors over CNN_{G23}, which is primarily coming from improvements in ice edge errors. This is consistent with the larger magnitude negative increments from CNN_{opt} in regions such as the Beaufort and Chukchi seas. This may then suggest that the positive summer

increments seen in the DA and CNN_{CNN+DA} corrections are needed to target local summer SIC errors. Regarding DA_{CNN+DA}, we can see that, in both hemispheres, these corrections are lower in magnitude than DA_{Z21} on average, which highlights how the initial correction from CNN_{CNN+DA} is removing a sizable component of the model error; leaving less error to correct with DA. This is particularly the case in the Antarctic, where the daily increments from DA_{CNN+DA} are very close to zero, suggesting that the CNN has effectively removed the systematic component of the model error in the Antarctic. Meanwhile in the Arctic, there are still residual systematic summertime errors associated with under-predicting the positive increments, which DA_{CNN+DA} needs to address.

A natural question then arises as to why DA_{Z21} is as effective, if not more effective, at reducing the model bias than CNN_{opt}, even though the increments from CNN_{opt} are considerably larger in magnitude. For this we turn to a comparison of the increments from each of the concentration categories (Figures S6 and S7 in Supporting Information S1). For CNN_{opt}, we find that the largest magnitude corrections are being made to the thinnest ice category, while on the other hand, DA_{Z21} makes sizable corrections to some of the thicker categories. This therefore means that CNN_{opt} needs to make larger corrections to achieve the same volume change as DA_{Z21}. Furthermore, the fact that CNN_{opt} is largely updating the thinnest ice category also explains the model shock seen in the regional Antarctic SIE time series for CNN_{opt} in Figure S2 in Supporting Information S1. In the Pacific sector in summer for example, the CNN is adding large extents of ice, which the model is then consistently removing over each 5-day interval. This now seems conceivable given that the new ice is very thin (5 cm thickness), and hence would be susceptible to completely melting if advected to grid cells with sufficiently warm SSTs. Further evidence to support this claim comes from the fact that this model shock behavior is significantly damped in the regional sea ice volume time series (Figure S8 in Supporting Information S1); which makes sense as the thinnest ice category will typically contribute less to the regional volume.

4. Discussion and Conclusions

There is currently much discourse centered around ML-based parameterizations and/or corrections within climate models, particularly in the context of how to achieve stable and unbiased simulations after implementation. Many studies have illustrated how to achieve stability within idealized models, where parameterizations are learned from high resolution simulations. For example, by swapping out neural networks for random forests (Watt-Meyer et al., 2021; Yuval & O'Gorman, 2020), or using online and/or reinforcement learning (Kurz et al., 2023; Rasp, 2020). In the context of DA-based approaches, previous studies have shown that iterative sequences of DA and ML can be used to learn the underlying dynamics of the numerical model to improve forecast capabilities in idealized systems (Bocquet et al., 2020; Brajard et al., 2020; Farchi et al., 2021, 2023).

In this study, we have built upon these previous DA-ML methodologies by exploring their application within a large-scale sea ice model used for climate simulations at GFDL. Subsequently, we have shown that a CNN model which has been trained purely offline to predict increments from a sea ice DA system (which assimilates real sea ice observations) can be used "out-of-the-box" to systematically reduce sea ice biases in a 5-year global ice-ocean simulation, without instabilities or drift. We have also introduced a data augmentation approach to optimize the offline-trained CNN, which significantly improves online generalization in both hemispheres; particularly in terms of reducing sea ice edge errors in summer. This augmentation approach is performed by iteratively generating new simulations in which corrections are applied from both the current iteration of the ML model, as well as DA. Each iteration of the augmentation procedure therefore provides a new training data set with which to refine the CNN weights from the previous iteration. While, in theory, this procedure could be repeated to convergence, we opted for $N = 3$ iterations in this study due to computational expense. It is likely however that continued iterations would yield further improvements, particularly in the Arctic summer, given that sizable gains were made between each of the three iterations here (see Figure S9 in Supporting Information S1). We hypothesize that the improvements from this augmentation procedure are a result of exposing the network to input variables which contain information about how the model trajectory evolves after implementing the CNN (as opposed to training purely offline where the inputs have no feedback with the CNN).

Interestingly, we find that, relative to the original DA experiment, the climatological sea ice biases associated with the simulation which uses this "optimized" network are actually modestly improved in the Antarctic (Figure 2). This is understandable when we consider that the target variable during each iteration of the network refinement step is no longer the increment from original DA experiment, but rather the sum of the increments from the two-step CNN + DA experiment (recall Figure 1b). Therefore, the model bias from the original DA

experiment should not be seen as the lower limit on what is achievable with the CNN. We can see this in Figure S10 in Supporting Information S1, where the bias of the simulation which applies this two-step CNN + DA procedure is indeed systematically lower than both the original DA experiment and the optimized CNN. As discussed in previous literature (Brajard et al., 2020; Farchi et al., 2021, 2023; Laloyaux et al., 2022), this combined DA-ML approach therefore leaves exciting avenues for future work relating to improved initial conditions for numerical prediction. For seasonal predictions with the GFDL SPEAR model for example, initial conditions for the ice and ocean (Y. Zhang et al., 2022; Lu et al., 2020) are based on DA via Ensemble Kalman filters. The Ensemble Kalman filter is not formally designed to correct for systematic model error, and so this two-step CNN + DA procedure could be a way to generate more accurate initial conditions (see Figure S10 in Supporting Information S1) with the CNN and DA fixing the systematic and random components of the errors, respectively. Indeed while this is similar to a weak constraint 4-D variational DA approach (Trémolet, 2007; Wergen, 1992; Zupanski, 1993), the CNN has computational advantages (once trained) in that it does not require the construction of an adjoint model (Bonavita & Laloyaux, 2020; Farchi et al., 2023; Laloyaux et al., 2022).

Further avenues for future work also include the use of the optimized CNN for making bias corrections to real-time seasonal sea ice forecasts, or sea ice projections on climate timescales. For seasonal prediction, the methodology would follow that which has been presented in this study, except the CNN would be applied within the fully coupled SPEAR model. This would however require several considerations. For example, addressing the issue of model shock seen in the Antarctic (which could potentially be reduced by increasing the frequency of the CNN corrections), and also assessing generalization of the CNN to the fully coupled model, which includes new interactive feedbacks with an atmospheric model. Looking also to longer term climate projections, G23 discussed this in the context of implementing the CNN as a sea ice model parameterization. This would then require further considerations of how to appropriately conserve mass, heat and salt when adding/removing sea ice from the ocean.

Data Availability Statement

All data from the 1982–2017 DA_{Z21} experiment (which were used to train the CNN_{G23} network) are available through a v.1.0.0 release of the software on Zenodo (Gregory, 2023a). This also includes the network weights and standardization statistics. The code to implement the CNN into the SIS2 sea ice model, as well as perform DA, are then available through a v.1.0.1 release on Zenodo (Gregory, 2023b). Network weights are also available after each iteration of the augmentation procedure. The baseline DA_{Z21} data and available code should enable to reader to reproduce the results of this paper. WG is pleased to assist in configuring any experiments which seek to replicate this work.

Acknowledgments
William Gregory, Mitchell Bushuk, Alistair Adcroft and Laure Zanna received M^2LInES 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^2LInES 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 Theresa Morrison and Feiyu Lu for their invaluable feedback on this article.

References

Adcroft, A., Anderson, W., Balaji, V., Blanton, C., Bushuk, M., Dufour, C. O., et al. (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. https://doi.org/10.1029/2019MS001726

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

Arcucci, R., Zhu, J., Hu, S., & Guo, Y.-K. (2021). Deep data assimilation: Integrating deep learning with data assimilation. *Applied Sciences*, *11*(3), 1114. https://doi.org/10.3390/app11031114

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. https://doi.org/10.5194/essd-8-165-2016

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*, *106*(C2), 2441–2463. https://doi.org/10.1029/1999JC000113

Bocquet, M., Brajard, J., Carrassi, A., & Bertino, L. (2020). Bayesian inference of chaotic dynamics by merging data assimilation, machine learning and expectation-maximization. arXiv preprint arXiv:2001.06270. https://doi.org/10.3934/fods.2020004

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

Brajard, J., Carrassi, A., Bocquet, M., & Bertino, L. (2020). Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: A case study with the Lorenz 96 model. *Journal of Computational Science*, *44*, 101171. https://doi.org/10.1016/j.jocs.2020.101171

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. https://doi.org/10.1098/rsta.2020.0086

Brenowitz, N. D., Beucler, T., Pritchard, M., & Bretherton, C. S. (2020). Interpreting and stabilizing machine-learning parametrizations of convection. *Journal of the Atmospheric Sciences*, *77*(12), 4357–4375. https://doi.org/10.1175/JAS-D-20-0082.1

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

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. 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). https://doi.org/10.1029/2022MS003309

Delworth, T. L., Cooke, W. F., Adcroft, A., Bushuk, M., Chen, J.-H., Dunne, K. A., et al. (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. https://doi.org/10.1029/2019MS001895

DiGirolamo, N., Parkinson, C., Cavalieri, D., Gloersen, P., & Zwally, H. (2022). *Sea ice concentrations from Nimbus-7 SMMR and DMSP SSM/I-SSMIS passive microwave data, version 2*. National Snow and Ice Data Center. https://doi.org/10.5067/MPYG15WAA4WX

Driscoll, S., Carrassi, A., Brajard, J., Bertino, L., Bocquet, M., & Olason, E. (2023). Parameter sensitivity analysis of a sea ice melt pond parametrisation and its emulation using neural networks. arXiv preprint arXiv:2304.05407. https://doi.org/10.48550/arXiv.2304.05407

Farchi, A., Chrust, M., Bocquet, M., Laloyaux, P., & Bonavita, M. (2023). Online model error correction with neural networks in the incremental 4D-Var framework. *Journal of Advances in Modeling Earth Systems*, *15*(9), e2022MS003474. https://doi.org/10.1029/2022MS003474

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. https://doi.org/10.1002/qj.4116

Finn, T., Durand, C., Farchi, A., Bocquet, M., Chen, Y., Carrassi, A., & Dansereau, V. (2023). Deep learning subgrid-scale parameterisations for short-term forecasting of sea-ice dynamics with a Maxwell Elasto-Brittle rheology. *The Cryosphere*, *17*(7), 2965–2991. https://doi.org/10.5194/tc-17-2965-2023

Frezat, H., Le Sommer, J., Fablet, R., Balarac, G., & Lguensat, R. (2022). A posteriori learning for quasi-geostrophic turbulence parametrization. *Journal of Advances in Modeling Earth Systems*, *14*(11), e2022MS003124. https://doi.org/10.1029/2022MS003124

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. https://doi.org/10.1029/2018GL078202

Goessling, H. F., Tietsche, S., Day, J. J., Hawkins, E., & Jung, T. (2016). Predictability of the arctic sea ice edge. *Geophysical Research Letters*, *43*(4), 1642–1650. https://doi.org/10.1002/2015GL067232

Gregory, W. (2023a). Sea ice DA-ML: April 11, 2023 release (version 1.0.0) [Dataset, Software]. Zenodo. https://doi.org/10.5281/zenodo.7818178

Gregory, W. (2023b). Sea ice DA-ML: December 11, 2023 release (version 1.0.1) [Software]. Zenodo. https://doi.org/10.5281/zenodo.10359842

Gregory, W., Bushuk, M., Adcroft, A., Zhang, Y., & Zanna, L. (2023). Deep learning of systematic sea ice model errors from data assimilation increments. *Journal of Advances in Modeling Earth Systems*, *15*(10), e2023MS003757. https://doi.org/10.1029/2023MS003757

Griffies, S. M., Biastoch, A., Böning, C., Bryan, F., Danabasoglu, G., Chassignet, E. P., et al. (2009). Coordinated ocean-ice reference experiments (COREs). *Ocean Modelling*, *26*(1–2), 1–46. https://doi.org/10.1016/j.ocemod.2008.08.007

He, Z., Brajard, J., Wang, Y., Wang, X., & Shen, Z. (2023). Improve dynamical climate prediction with machine learning. *ESS Open Archive*. https://doi.org/10.22541/essoar.167898499.91486328/v1

Held, I., Guo, H., Adcroft, A., Dunne, J., Horowitz, L., Krasting, J., et al. (2019). Structure and performance of GFDL's CM4. 0 climate model. *Journal of Advances in Modeling Earth Systems*, *11*(11), 3691–3727. https://doi.org/10.1029/2019MS001829

Kurz, M., Offenhäuser, P., & Beck, A. (2023). Deep reinforcement learning for turbulence modeling in large eddy simulations. *International Journal of Heat and Fluid Flow*, *99*, 109094. https://doi.org/10.1016/j.ijheatfluidflow.2022.109094

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*, *14*(6), e2022MS003016. https://doi.org/10.1029/2022MS003016

Lu, F., Harrison, M. J., Rosati, A., Delworth, T. L., Yang, X., Cooke, W. F., et al. (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. https://doi.org/10.1029/2020MS002149

Meier, W. N., & Stewart, J. S. (2023). *Arctic and Antarctic regional masks for sea ice and related data products, version 1*. National Snow and Ice Data Center. https://doi.org/10.5067/CYW3O8ZUNIWC

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. https://doi.org/10.1063/5.0091282

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. https://doi.org/10.1029/2018MS001351

Ott, J., Pritchard, M., Best, N., Linstead, E., Curcic, M., & Baldi, P. (2020). A Fortran-Keras deep learning bridge for scientific computing. *Scientific Programming, 2020*, *2020*, 1–13. https://doi.org/10.1155/2020/8888811

Rasp, S. (2020). Coupled online learning as a way to tackle instabilities and biases in neural network parameterizations: General algorithms and Lorenz 96 case study (v1.0). *Geoscientific Model Development*, *13*(5), 2185–2196. https://doi.org/10.5194/gmd-13-2185-2020

Rasp, S., Pritchard, M. S., & Gentine, P. (2018). Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(39), 9684–9689. 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. https://doi.org/10.1175/2007JCLI1824.1

Ross, A., Li, Z., Perezhogin, P., Fernandez-Granda, C., & Zanna, L. (2023). Benchmarking of machine learning ocean subgrid parameterizations in an idealized model. *Journal of Advances in Modeling Earth Systems*, *15*(1), e2022MS003258. https://doi.org/10.1029/2022MS003258

Sane, A., Reichl, B. G., Adcroft, A., & Zanna, L. (2023). Parameterizing vertical mixing coefficients in the ocean surface boundary layer using neural networks. *Journal of Advances in Modeling Earth Systems*, *15*(10), e2023MS003890. https://doi.org/10.1029/2023MS003890

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

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

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

Wergen, W. (1992). The effect of model errors in variational assimilation. *Tellus*, *44*(4), 297–313. 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. 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. https://doi.org/10.1029/2020GL088376

Zhang, C., Perezhogin, P., Gultekin, C., Adcroft, A., Fernandez-Granda, C., & Zannaaure, L. Z. (2023). Implementation and evaluation of a machine learned mesoscale eddy parameterization into a numerical ocean circulation model. *Journal of Advances in Modeling Earth Systems*, *15*(10), e2023MS003697. https://doi.org/10.1029/2023MS003697

Zhang, Y., Bushuk, M., Winton, M., Hurlin, B., Delworth, T., Harrison, M., et al. (2022). Subseasonal-to-seasonal Arctic sea ice forecast skill improvement from sea ice concentration assimilation. *Journal of Climate*, *35*(13), 1–48. 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. https://doi.org/10.1175/JCLI-D-20-0469.1

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