Benchmarking ML-Based Antarctic Sea Ice Forecasting in a Data-Rich Setting

Yuchen Li
Stanford University
yucli@stanford.edu

Earle Wilson*
Dept. of Earth System Science
Stanford University
earlew@stanford.edu

Zachary Kaufman*
Dept. of Earth System Science
Stanford University
zack_kaufman@stanford.edu

Abstract

Forecasting sea ice concentration on seasonal timescales is important for both climate science and polar operations, but machine learning methods are limited by the small size of the observational record. To address this, we train U-Net models on a large ensemble of climate model simulations (CESM-LE) to evaluate the effects of physical inputs, dataset size, and simulation-based pretraining. We find that including atmospheric and oceanic variables—especially sea level pressure—marginally improves skill in predicting summer sea ice, though not in ways easily linked to known climate modes. Increasing training data volume improves performance in difficult regimes, but yields diminishing returns. Pretraining on CESM-LE followed by finetuning on observational data improves winter predictions but degrades performance in summer, suggesting persistent biases inherited from the simulation. These results caution against naïve use of simulation data for pretraining and highlight the need for more robust transfer strategies.

1. Introduction

Sea ice is a seasonally-varying layer of ice that is ubiquitous in Earth's polar oceans. Understanding the dynamics and evolution of sea ice is important to understanding the broader polar climate, because it mediates transfers of heat, moisture, and momentum between the atmosphere and ocean. Additionally, sea ice is important for human activities in polar regions such as shipping, scientific research, fishing, and tourism.

Forecasting the spatial extent of sea ice with seasonal (i.e., 3 to 6 months) lead times has been a problem of both scientific and operational interest [2, 9]. Similar to weather forecasting, this can be viewed as an initial value problem that can be solved by constraining some initial state estimation via data assimilation, then stepping forward that state according to some dynamics. Traditionally, this is done by solving the equations of fluid flow and thermodynamics on a discrete grid (exactly analogous to numerical weather prediction) [10, 3].

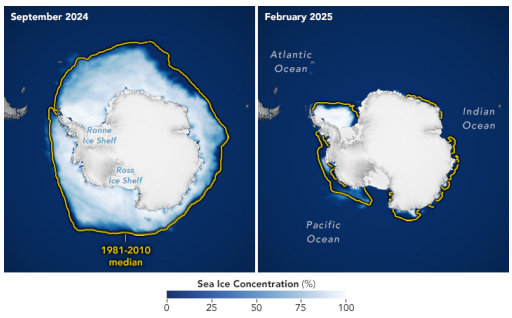

Figure 1. Antarctic sea ice in September (maximum extent month) and February (minimum extent month).

In recent years, numerous groups have shown that it is viable to train neural networks to forecast maps of sea ice concentration of comparable or even better skill than traditional physics-based forecasting methods [1, 16, 12, 19, 18]. The reasons for this are twofold: 1) operational prediction of sea ice, like weather, uses data assimilation to constrain initial conditions. However, our ability to do data assimilation is constrained by limited continuous observation in polar regions and in the subsurface ocean; 2) sea ice involves multiscale physics that must be represented heuristically in physical models.

However, these machine learning methods are often trained to predict maps of sea ice concentration obtained from the observational satellite record (1978 to present), which when sampled at monthly frequency amounts to $O(500)$ data points. The limited size of the dataset presents several challenges. First, it is possible that models capable of emulating the complex dynamics of this system are data-constrained, do not generalize well to extremes, or are prone to overfitting. Additionally, conducting ablation or scaling tests to systematically assess model performance is difficult on small test sets restricted to the observational dataset.

*Project mentors not enrolled in CS231N

To address these challenges, we relax the constraint of using purely observational data, instead opting to use an ensemble of simulations derived from a physics-based climate model. The motivation is therefore *not* to build the best-performing sea ice forecasting model that can emulate sea ice in the real world, but rather to probe the design space, capabilities, and limitations of ML-driven methods for sea ice forecasting in a data-rich setting. Additionally, using simulation data gives the opportunity to study whether or not it is beneficial to pretrain models on simulation data in order to better emulate real-world sea ice dynamics.

2. Related work

Andersson *et al*. [1] developed IceNet, a U-Net model that skillfully predicts the next six months of spatial coverage of Arctic sea ice given a diverse set of physical inputs, such as sea ice concentration, sea level pressure, sea surface temperature, etc., at various lags. Notably, they showed that their model was, on average, more skillful than a state-of-the-art physics based model, SEAS5.

Yang *et al*. [18] follow the framework in [1] to predict Antarctic sea ice on seasonal timescales. They find qualitatively similar results. In particular, in both [1] and [18], an input-permutation ablation test (i.e., where inputs of the same physical type are shuffled in the time dimension) showed that most physical inputs, even when permuted, did not significantly harm model performance. This suggests that many of the inputs are perhaps not necessary for recovering model performance. More recently, various groups have augmented the vanilla convolutional U-Net with spatiotemporal attention [13, 16] and neural ODEs [12], finding slight performance improvements.

Uebbing *et al*. [15] perform a feature ablation analysis on the original IceNet model [1] and find that the performance of models trained *only on past sea ice concentration* are within the epistemic uncertainty bounds of the performance of the the original model.

In the present work, we analyze the performance of convolutional models trained to forecast Antarctic sea ice up to six month lead times, similar to the setup in [1] and [18]. However, unlike previous studies which are trained on the observed sea ice, we train our models to predict simulated sea ice dynamics (CESM Large Ensemble, or CESM-LE). CESM-LE serves as a data-rich environment within which we can perform a more robust set of ablation and scaling experiments. Furthermore, we follow [1] in evaluating the performance of models pretrained on simulation data on the task of forecasting true (observed) sea ice. While [1] found a marginal improvement in pretrained models for forecasting Arctic sea ice, pretrained models have to our knowledge not yet been evaluated for forecasting Antarctic sea ice, which is typically represented more poorly by physical climate models [9, 3].

3. Dataset and methods

3.1. CESM Large Ensemble

The Community Earth System Model Large Ensemble (CESM-LE) is a set of climate simulations generated with the CESM, a state-of-the-art Earth system model [7]. Each ensemble member[1] is subject to identical external radiative forcings (e.g., historical greenhouse gas emissions) but initialized with roundoff error-level perturbations to the atmospheric state. These perturbations grow chaotically, producing divergent yet equally plausible climate trajectories that sample internal variability around the forced mean climate. Importantly, there is no assimilation of observed data, so these simulations, unlike reanalysis products, are not constrained to the *particular* realization of observed climate variability. The ensemble spans over a century at relatively high temporal (daily to monthly) and spatial ($\sim 1^\circ$) resolution and includes a comprehensive set of physical variables across atmosphere, ocean, land, and sea ice components.

Gridded data from CESM-LE are of the form $X \in \mathbb{R}^{T \times K \times N \times W \times H}$ where T is the temporal dimension, K is the physical variable dimension, N is the ensemble member, and W and H are the spatial dimensions. For our purposes, we regrid the data to a stereographic south polar projection (such that the South Pole is centered) with spatial dimensions $H \times W = 80 \times 80$. This spatial resolution is chosen such that it approximately retains the 1° resolution of the native model output and may be evenly divided in two for downsampling operations.

For each ensemble member, we normalize sea ice concentration by subtracting out the mean for each grid point (no additional normalization is done because sea ice concentration is already within the range of 0 and 1). For other inputs, we apply min-max normalization so that training inputs are within the range $[0, 1]$. Finally, we remove the forced climate trend by removing the quadratic least squares fit from each grid point.

3.2. Observational data

Observational sea ice data (monthly averaged from 1979 through 2024) is obtained from NSIDC and other variables are obtained from the ERA5 reanalysis. The former is derived from passive microwave satellite measurements, while the latter is a reanalysis product generated by data assimilation from varied remote sensing and in-situ sources. These gridded data are mapped to the same 80×80 polar grid as the processed CESM-LE data.

[1] Due to overloading of this term, we will explicitly distinguish between ensemble members in CESM (which correspond to distinct climate simulations) and U-Net ensemble members (which correspond to different realizations of our model trained with different random initializations and minibatches).

Name	Configuration	C
`input1`	Past 12 months of SIC, land mask, sin and cos of month index	15
`input2a`	input1 + past 6 months of SST	21
`input2b`	input1 + past 6 months of SLP	21
`input2c`	input1 + past 6 months of Z_{500}	21
`input3`	input1 + past 6 months of SST, SLP, and Z_{500}	33

Table 1. Experiment 1 (inputs) configurations. SIC = sea ice concentration; SST = sea surface temperature; SLP = sea level pressure; Z_{500} = geopotential height at 500 mb. C is the total number of input channels.

3.3. Models

We use a relatively unmodified U-Net architecture, similar to what is used in [1] and [18]. The U-Net has a standard encoder-bottleneck-decoder structure with three downsampling and upsampling layers (see Supplementary Info A.1). The model is trained to forecast maps of deviations from the time-average sea ice concentration (in climate science this is referred to as the "anomaly from climatology"). This is motivated by the fact that by far the strongest signal in the raw sea ice data is the seasonal cycle (see Figure 1), and therefore the skillfulness of predictions is always assessed relative to the background seasonal cycle.

Our model does not explicitly treat temporal dynamics and takes inputs with shape $\mathbb{R}^{N\times C\times H\times W}$, where C includes different physical inputs at different lag times and N is the batch dimension. The output is of shape $\mathbb{R}^{N\times 6\times H\times W}$ since the model forecasts one map for each month, up to six months. The reason for using a simple architecture is that: 1) U-Nets and their variants have been used in most preexisting studies in this field; 2) more advanced architectures seem to provide only marginal gains in improvement; 3) the U-Net we adopt is relatively lightweight (8M parameters) and easy to train.

4. Experiments

4.1. Experiment 1: Physical inputs

We test the benefit of adding additional oceanic and atmospheric variables to the input channels. The input configurations are described in Table 1. We use the same training settings to ensure consistency among the configurations. The models are trained using Adam optimizer with a learning rate of 0.001 and a batch size of 64. We use an area-weighted MSE loss that also accounts for seasonality in sea ice area (see Supplementary Info A.2). We use an 8-ensemble member subset of CESM-LE for training ($n = 15648$ samples), 2 ensemble members for validation, and 4 ensemble members for testing. Models are trained for 10 epochs (based on experimentation, further training results in overfitting to the training dataset). To assess robustness to initialization, 5 models from unique weight initializations are trained to generate an ensemble of U-Nets.

The accuracy of model predictions is assessed by the spatial anomaly correlation coefficient (ACC). Given a single prediction map $\hat{Y} \in \mathbb{R}^S$ where $S = H \times W$ is the flattened spatial dimension and true label Y, the anomaly correlation coefficient is defined as the Pearson correlation in the spatial dimension:

$$\begin{aligned}\mathrm{ACC}(\hat{Y}, Y) &= \frac{\sum_{i=1}^{S}(Y_i - \overline{Y})(\hat{Y}_i - \overline{\hat{Y}})}{\sqrt{\sum_{i=1}^{S}(Y_i - \overline{Y})^2 \sum_{i=1}^{S}(\hat{Y}_i - \overline{\hat{Y}})^2}} \\ &= r(\hat{Y}, Y)\end{aligned}$$

For this experiment, we calculate an ACC score for each model prediction. We aggregate these statistics according to the target month and lead time. We then conduct a simple bootstrap to assess statistical significance in differences between ACC scores among different input configurations (Supplementary Info B.1).

First, we discuss the characteristics of sea ice predictability that are general across input configuration. Figure 2 shows that the `input1` baseline, trained to predict future sea ice only with past sea ice, is skillful at short lead time across the year. This skill extends to longer lead times in the winter due to persistence of anomalies through the ice growth season. The model is generally unskillful for predicting fall and summer sea ice at long lead time, consistent with the fact that the ice melt season is more strongly driven by synoptic atmospheric processes (i.e., weather that is unpredictable months in advance) and that oceanic memory is cut off due to freshening of the upper mixed layer [8].

Adding sea surface temperature, sea level pressure, and 500 mb geopotential height (`input3`) generally improves model skill relative to the model trained using only sea ice as input (`input1`). The improvement is most pronounced (order 10%) for predicting summer months (JFM) at long lead time. This is promising, as predicting summertime conditions through the melt season is the hardest time of year to make skillful predictions [3, 6, 11]. When the three inputs are added separately, it is revealed that the improvement can be attributed primarily to addition of sea level pressure.

In contrast, addition of sea surface temperature alone (`input2a`) indeed *harmed* the model performance for predicting summertime conditions at 5 to 6 months lead time. We surmise that this may be due to the tendency for under-ice SST anomalies to be strongly out-of-distribution. The typical range of SST under sea ice is constrained tightly to the freezing point of seawater, so small e.g., $O(0.01^\circ\,\mathrm{C})$ anomalies can be strongly amplified even under min-max normalization (see Supplementary Info B.2 and Figure S2).

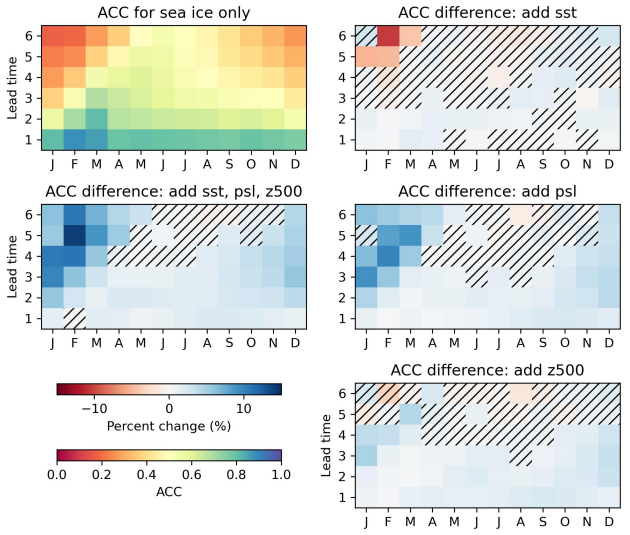

Figure 2. Experiment 1 (inputs). Top left: average ACC for sea ice-only configuration (`input1`) as a function of target month and lead time (e.g., bottom left box corresponds to predicting January sea ice at a lead time of 1 month, so inputs are provided up to the previous December). Other plots: changes in ACC, expressed as percent differences. Hatched boxes are not statistically significant at the $p = 0.05$ level.

Next, we ask the question: are instances for which `input3` model outperforms the `input1` model associated with a physically consistent signal in the inputs?

To answer this question, we use PCA to extract the dominant modes of variability in the additional physical inputs (SST, SLP, and Z_{500}) provided to `input3`. For each variable, this yields the leading eigenvectors of the covariance matrix; then, for each instance in time we compute the projection of the true data onto the leading eigenvectors to get the principal component timeseries. We plot joint distributions of the leading components with the ensemble-mean ACC difference between `input3` and `input1` configurations. Since the greatest improvements in skill are for predicting January through March at 4-6 month lead time, we select only instances where the model prediction starts in September (thus reaching March at lead time 6). In Figure 3, we show results for the first principal component in the sea level pressure, which is chosen due to the fact that sea level pressure seemed to have the biggest role in improving model predictions (see Figure 2). Furthermore, the first principle component of sea level pressure corresponds to the Southern Annular Mode (SAM), a mode of climate variability known to have a driving effect on sea ice variability [4, 14] and accounts for 24.6% of the variance in the sea level pressure data.

However, it is clear from Figure 3 that there is no relationship between the magnitude of SAM and improvements

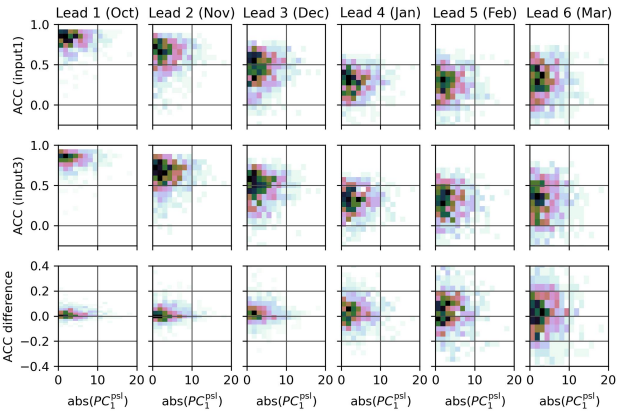

Figure 3. Joint distributions of ACC and the absolute value of the first principle component in sea level pressure for predictions initialized in September. The first two rows show the U-Net ensemble-mean ACC for each input configuration, while the bottom row shows $\mathrm{ACC}_{\mathrm{input3}} - \mathrm{ACC}_{\mathrm{input1}}$.

in the `input3` model relative to the `input1` model. We surmise that the mechanisms leading to additional summertime predictability in the `input3` model cannot be captured in this relatively simple regression onto the first principle component. Further work towards this end may benefit from saliency or attribution maps.

4.2. Experiment 2: Dataset scaling

Motivated by the small size of the observational sea ice dataset, we conduct a data scaling experiment using CESM-LE data to test the effect of training dataset size on generalization skill. Training datasets are prepared with 1, 4, 16, and 64 CESM ensemble members (one CESM ensemble member corresponds to $n = 1956$ data points). For simplicity, we adopt the `input1` configuration from Experiment 1, though we expect these results to generalize to other input configurations as well. We then train the same U-Net model architecture using the a common set of optimization settings (identical to Experiment 1). Training is continued until early stopping is triggered with a patience of 5 epochs. We use the same held out 2 CESM ensemble members for validation and 5 CESM ensemble members for testing.

Figure 4 shows the average spatial ACC as a function of lead time and target month for each of the dataset sizes. First, we observe considerable improvements in ACC for forecasting summer sea ice months at long lead time, especially when going from 1 to 4 and 4 to 16 CESM ensemble members. However, these improvements diminish when we further expand the training dataset to include 64 CESM ensemble members. Another important feature is that targets associated with greater overall predictability—for example one month lead time for the entire year and winter

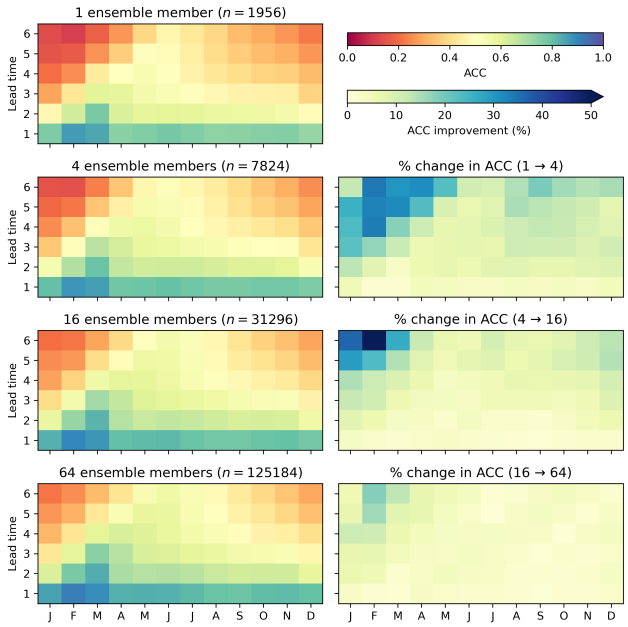

Figure 4. Experiment 2 (data scaling). Left subplots show the ACC as a function of lead time and target month for each dataset size. Right subplots show percent changes in ACC from one scaling stage to the next.

months (JJA)—do not show significant improvement when the dataset size is increased. This pattern is consistent with the times during which temporal persistence of sea ice anomalies is strong and therefore predictability is relatively easy to learn. The nature of these results is perhaps expected, but it is clarifying to see that the trends in model

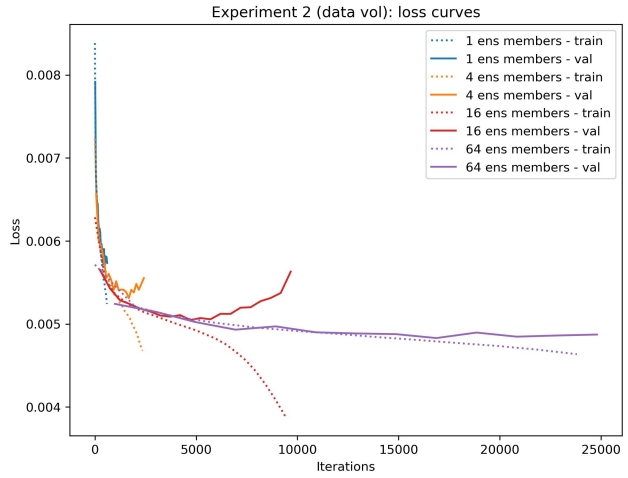

Figure 5. Loss curves for the data scaling experiment. Dashed lines correspond to training loss and solid lines correspond to validation loss.

improvement with dataset scaling indeed correspond with underlying regimes of predictability (or lack thereof). Additionally, we note that the magnitude of improvements due to data scaling alone is greater than improvements from adding additional physical inputs with dataset size fixed (Figure 2). Therefore, while not explicitly tested here, we speculate that the results of this experiment would not be significantly different if another input configuration (i.e., additional physical inputs) is used.

Another perspective comes from inspecting the loss curves for each of the runs (Figure 5), which clearly shows that the model is more robust to overfitting when trained on a larger dataset. This suggests that models trained on the observational sea ice dataset may be significantly constrained by data sparsity, at least with respect to standard supervised learning optimization techniques as used here.

4.3. Experiment 3: Does pretraining help?

We test the efficacy of pretraining on the CESM dataset before finetuning to the observational dataset. The objective, like before, is to predict sea ice concentration anomalies from the mean. Here we use the mean of the finetuning (observational) dataset, so that sea ice inputs and targets are defined by $\mathrm{SIC}' = \mathrm{SIC} - \overline{\mathrm{SIC}}_{\mathrm{obs,train}}$ (SIC = sea ice concentration and $\overline{\{\cdot\}}$ denotes temporal averaging). Note that the possibility of distribution shift possibly complicates the interpretation of this definition; we discuss this fact later.

We use initial weights from the model trained on the largest CESM dataset (64 ensemble members) from Experiment 2. We then finetune all model weights (encoder and decoder) on years 1979–2011 and conduct an extensive hyperparameter sweep over a validation dataset of 2012–2015

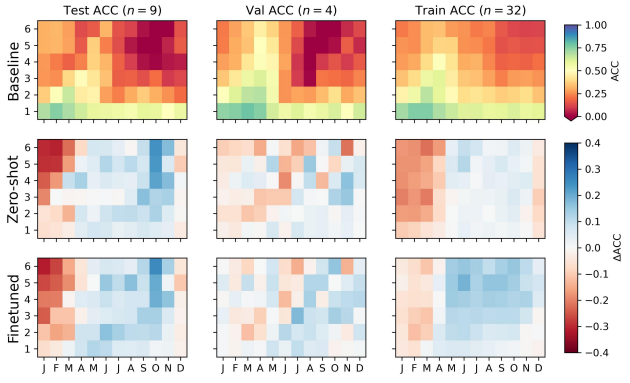

Figure 6. Experiment 3 (finetuning). The top row shows ACC scorecards for the baseline model trained only on observations. The middle row shows the ACC difference for the zero-shot 64-ensemble member model. The bottom row shows the ACC difference for the finetuned model. For each data split, the number of years in that subset is denoted as $(n = \ldots)$. Thus, each cell represents an average over n samples.

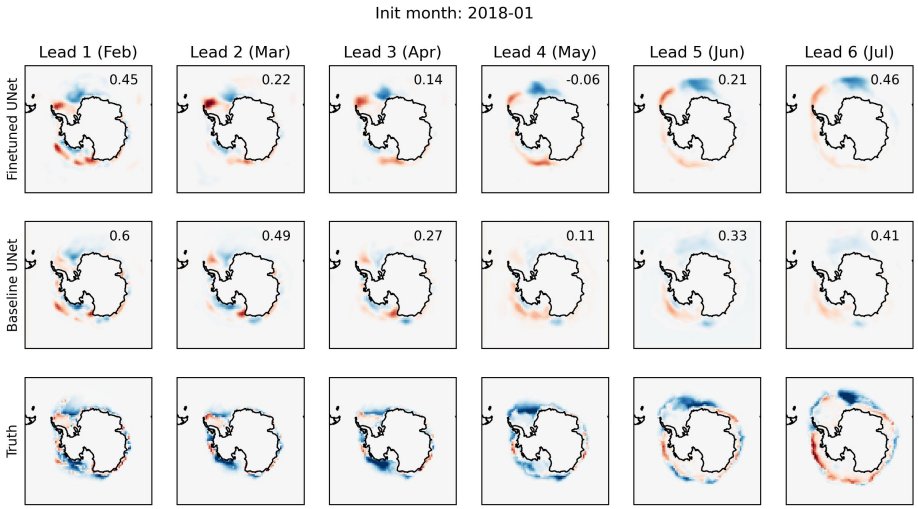

Figure 7. Experiment 3 (finetuning). Sample prediction from the finetuned and baseline models. Blue regions represent negative anomalies while red regions represent positive anomalies. The ACC for each prediction is printed on the top right. This specific sample corresponds to inputs up to January 2018 and predictions up to July (winter). As is exemplified here, the finetuned model typically has more confident predictions near the ice edge for the ice growth season. By contrast, the predictions of the baseline model are more diffuse and smaller in magnitude.

of the observational dataset. For each hyperparameter setting, we train until overfitting using early stopping with a patience of 5 epochs and consider the best validation loss. Interestingly, we found that the models with the best validation loss achieved this loss within one or two epochs of finetuning with a relatively large initial learning rate. On the other hand, models finetuned using a smaller learning rate, while more resistant to overfitting, did not generalize as well over the validation set. The results are computed for the best model from the sweep over the testing dataset which spans 2016–2024. More details of the training and hyperparameter sweep are provided in Supplemental Information C.1. We compare the finetuned model to two baselines: 1) a model trained only on the observational dataset (from here we will simply refer to this as the baseline model); 2) zero-shot evaluation of the pretrained model on observations.

While one might expect that the finetuned model exhibit unequivocal improvement relative to the baseline due to the small size of the observational dataset, Figure 6 shows that this is indeed not the case. First, we note that the performance of the baseline model on observational dataset is considerably worse than even the model trained on a single CESM ensemble member (Figure 4). This discrepancy is particularly pronounced in wintertime sea ice prediction. Next, we note that the finetuned model and the zero-shot model both perform better over wintertime sea ice prediction, but worse for summertime sea ice prediction. Interestingly, this pattern is consistent across all lead times.

The similarity in performance between the finetuned and zero-shot models indicates that the two models are not far in parameter space. This is consistent with our earlier observation that in the hyperparameter sweep, the models with the best validation loss tended to achieve it within one or two epochs before starting to overfit. Furthermore, the "better in winter, worse in winter" pattern is present even in the zero-shot model evaluated on the training dataset, which suggests that this is a bias that the model carries over from pretraining.

Inspecting the model predictions shows that the finetuned model is often more confident, generating predictions with larger magnitude anomalies and sharper edges than the baseline model (Figure 7). This typically helps the finetuned model in predicting the ice growth season through fall and into winter, as anomalies are generally persistent, pronounced, and localized near the ice edge. By contrast, Figure S6 exemplifies the fact that both finetuned and baseline models essentially exhibit no skill in predicting the melt season through spring and summer.

Next, we ask: is the seasonal pattern in differences in performance between models trained with CESM data and purely observational data attributable to mean state biases between CESM and observational sea ice?

To address this question, we recompute the normalized observational data using mean statistics from CESM-LE; that is, sea ice anomaly inputs and targets are redefined as $\mathrm{SIC}' = \mathrm{SIC} - \overline{\mathrm{SIC}}_{\mathrm{CESM,pretrain}}$. This is akin to pretending that observations is yet another ensemble member from CESM. We then finetune the same model trained on 64 ensemble members of CESM to predict this new ob-

jective. When evaluating the model, we convert the predicted anomalies back to anomalies relative to the observational dataset by adding the correction $\overline{\text{SIC}}_{\text{CESM,pretrain}} - \overline{\text{SIC}}_{\text{obs,train}}$. The results from this experiment are shown in Figure S5. We find that adjusting inputs and targets to account for the mean state bias in fact *harms* the finetuned model performance compared to the standard normalization where the observational mean is used. We posit the following explanation. Shifting the mean will generally result in more persistent anomalies, especially at the ice edge, that correspond to the time-averaged difference between $\overline{\text{SIC}}_{\text{CESM,pretrain}}$ and $\overline{\text{SIC}}_{\text{obs,train}}$ rather than meaningful anomalies from the mean. Therefore, the pretrained model, which has learned to dampen persistent anomalies over some decorrelation timescale, will encounter during finetuning anomalies that are more persistent than what it saw during pretraining.

Furthermore, we interpret this finding to indicate that the pretrained model carries over not only mean state biases, but more importantly biases in the dynamics of anomalies from the mean. The exact characteristics of this bias are left for future work. Overall, our findings in this experiment suggest that pretraining on possibly biased simulation data may yield adverse results, even if the models perform much better within the pretraining dataset.

5. Discussion

Our analysis of ML-based sea ice forecasting models in a relatively data rich setting has revealed the following new insights:

1. Addition of additional atmospheric and oceanic input variables leads to order 10% gains in improvement for predicting summer sea ice. This is mostly attributable to sea level pressure (SLP), but we find no relationship between the dominant modes of SLP variability and improvements in model skill. Addition of SST alone can harm model performance.

2. Scaling the size of the training dataset in CESM-LE results in significant improvements in difficult-to-predict regimes such as summer at long lead time, where anomaly persistence is a poor baseline mechanism of predictability. Conversely, essentially no additional skill is extracted in regimes where persistence is the dominant mode of predictability, such as one month lead times across the whole year as well as winter months at longer lead times. We find that models trained on larger datasets are less prone to overfitting.

3. Finetuning a model pretrained on CESM data to predict observational data yields improvement relative to a non-pretrained baseline for predicting April through November, but worse performance for summer months January through March. This pattern is present for both zero-shot and finetuned models, suggesting that biases learned from CESM persist through the finetuning process. We find that redefining the sea ice objective to be relative to CESM means does not improve finetuned model performance, suggesting that this bias lies in the dynamic mechanisms of predictability and not in differences mean sea ice.

Perhaps the most surprising and important result is that the pretrained model does not improve upon the data-limited baseline model in all target months. The cautionary tale illustrated here is that the pretraining dataset may result in learned biases that are not able to be corrected during finetuning, even though added data richness significantly improves performance within the pretraining dataset. A natural extension of the finetuning experiment is to analyze the effectiveness of pretraining on other simulation datasets that might better resemble real-world sea ice dynamics.

We will conclude with some reflections on the future prospects of using machine learning for emulation or forecasting of sea ice. We focus in this study on the characteristics of deterministic models trained to predict monthly-averaged statistics. One shortcoming of this approach is that we have made no serious attempt to quantify the inherent uncertainty in subseasonal to seasonal climate dynamics. While neural ensembling (which we perform in Exp. 1) is a potential avenue for this uncertainty quantification, it is not yet clear that the distribution over a neural ensemble should resemble the expected distribution over uncertainty due to sensitivity of the earth system to initial conditions. We believe that it will be valuable for future work to focus on the calibration of e.g., probabilistic or (conditional) generative models to the emulation of sea ice dynamics ([5, 17]).

Another shortcoming that we do not address here is that our model predicts monthly-averaged quantities and is therefore limited to learning dynamics of processes that occur with timescales longer than one month. This is a reasonable baseline for sea ice forecasting on 6 month timescales, since sea ice is a relatively slowly-evolving part of the earth system. On the other hand, studies like [13] and [12] focus only on short-term sea ice forecasting limited to a lead time of seven days. However, to predict across weekly to subseasonal, seasonal, and annual timescales, we anticipate that future ML models of the sea ice system will jointly model sea ice, ocean, and atmospheric dynamics (as is done in traditional physics-based earth system models), in which case daily or subdaily timestepping will need to be used. We foresee that effectively supervising the learning of both short timescale and long timescale dynamics in such a model will be a challenge.

6. Contributions

Y.L. designed the experiments, wrote the code, trained the neural networks, performed the evaluation, analyzed the results, and wrote this paper. E.W. suggested the analysis of correcting for mean state bias in Exp. 3. Both E.W. and Z.K. (not enrolled in CS231N) provided mentorship through meetings throughout this past academic year and helped with project conceptualization.

This project is a continuation of an ongoing research project (involving the three listed authors) that initially began in July 2024. *The main results presented in all three experiments were newly obtained during this academic quarter*. However, most of the codebase used to run these experiments (except for Experiment 3) was developed prior to this quarter. Moreover, we conducted some initial experiments prior to this quarter that informed the particular ones shown here.

7. Acknowledgements

Y.L. acknowledges funding from [fill this in here]. We thank the Sherlock team and Stanford Research Computing for providing the computational resources for training the models in this project. Y.L. thanks helpful feedback from Polar Ocean Dynamics group meetings and from discussions with the climate-ML community in the Earth System Science department.

References

[1] T. R. Andersson, J. S. Hosking, M. Pérez-Ortiz, B. Paige, A. Elliott, C. Russell, S. Law, D. C. Jones, J. Wilkinson, T. Phillips, J. Byrne, S. Tietsche, B. B. Sarojini, E. Blanchard-Wrigglesworth, Y. Aksenov, R. Downie, and E. Shuckburgh. Seasonal Arctic sea ice forecasting with probabilistic deep learning. *Nature Communications*, 12(1):5124, Aug. 2021. Number: 1 Publisher: Nature Publishing Group.

[2] M. Bushuk, S. Ali, D. A. Bailey, Q. Bao, L. Batté, U. S. Bhatt, E. Blanchard-Wrigglesworth, E. Blockley, G. Cawley, J. Chi, F. Counillon, P. G. Coulombe, R. I. Cullather, F. X. Diebold, A. Dirkson, E. Exarchou, M. Göbel, W. Gregory, V. Guemas, L. Hamilton, B. He, S. Horvath, M. Ionita, J. E. Kay, E. Kim, N. Kimura, D. Kondrashov, Z. M. Labe, W. Lee, Y. J. Lee, C. Li, X. Li, Y. Lin, Y. Liu, W. Maslowski, F. Massonnet, W. N. Meier, W. J. Merryfield, H. Myint, J. C. A. Navarro, A. Petty, F. Qiao, D. Schröder, A. Schweiger, Q. Shu, M. Sigmond, M. Steele, J. Stroeve, N. Sun, S. Tietsche, M. Tsamados, K. Wang, J. Wang, W. Wang, Y. Wang, Y. Wang, J. Williams, Q. Yang, X. Yuan, J. Zhang, and Y. Zhang. Predicting September Arctic Sea Ice: A Multi-Model Seasonal Skill Comparison. *Bulletin of the American Meteorological Society*, -1(aop), Apr. 2024. Publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society.

[3] M. Bushuk, M. Winton, F. A. Haumann, T. Delworth, F. Lu, Y. Zhang, L. Jia, L. Zhang, W. Cooke, M. Harrison, B. Hurlin, N. C. Johnson, S. B. Kapnick, C. McHugh, H. Murakami, A. Rosati, K.-C. Tseng, A. T. Wittenberg, X. Yang, and F. Zeng. Seasonal Prediction and Predictability of Regional Antarctic Sea Ice. *Journal of Climate*, 34(15):6207–6233, Aug. 2021. Publisher: American Meteorological Society Section: Journal of Climate.

[4] E. W. Doddridge and J. Marshall. Modulation of the seasonal cycle of antarctic sea ice extent related to the southern annular mode. *Geophysical Research Letters*, 44(19):9761–9768, 2017.

[5] T. S. Finn, C. Durand, A. Farchi, M. Bocquet, and J. Brajard. Towards diffusion models for large-scale sea-ice modelling, 2024.

[6] M. M. Holland, E. Blanchard-Wrigglesworth, J. Kay, and S. Vavrus. Initial-value predictability of Antarctic sea ice in the Community Climate System Model 3. *Geophysical Research Letters*, 40(10):2121–2124, 2013. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/grl.50410.

[7] J. E. Kay, C. Deser, A. Phillips, A. Mai, C. Hannay, G. Strand, J. M. Arblaster, S. C. Bates, G. Danabasoglu, J. Edwards, M. Holland, P. Kushner, J.-F. Lamarque, D. Lawrence, K. Lindsay, A. Middleton, E. Munoz, R. Neale, K. Oleson, L. Polvani, and M. Vertenstein. The community earth system model (cesm) large ensemble project: A community resource for studying climate change in the presence of internal climate variability. *Bulletin of the American Meteorological Society*, 96(8):1333 – 1349, 2015.

[8] S. Libera, W. Hobbs, A. Klocker, A. Meyer, and R. Matear. Ocean-Sea Ice Processes and Their Role in Multi-Month Predictability of Antarctic Sea Ice. *Geophysical Research Letters*, 49(8):e2021GL097047, 2022. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2021GL097047.

[9] F. Massonnet, S. Barreira, A. Barthélemy, R. Bilbao, E. Blanchard-Wrigglesworth, E. Blockley, D. H. Bromwich, M. Bushuk, X. Dong, H. F. Goessling, W. Hobbs, D. Iovino, W.-S. Lee, C. Li, W. N. Meier, W. J. Merryfield, E. Moreno-Chamarro, Y. Morioka, X. Li, B. Niraula, A. Petty, A. Sanna, M. Scilingo, Q. Shu, M. Sigmond, N. Sun, S. Tietsche, X. Wu, Q. Yang, and X. Yuan. SIPN South: six years of coordinated seasonal Antarctic sea ice predictions. *Frontiers in Marine Science*, 10, May 2023. Publisher: Frontiers.

[10] D. Notz and C. M. Bitz. Sea ice in Earth system models. In *Sea Ice*, pages 304–325. John Wiley & Sons, Ltd, 2017. Section: 12 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118778371.ch12.

[11] A. C. Ordoñez, C. M. Bitz, and E. Blanchard-Wrigglesworth. Processes Controlling Arctic and Antarctic Sea Ice Predictability in the Community Earth System Model. Dec. 2018. Section: Journal of Climate.

[12] J. Park, S. Hong, Y. Cho, and J.-J. Jeon. Unicorn: U-Net for Sea Ice Forecasting with Convolutional Neural Ordinary Differential Equations, Sept. 2024. arXiv:2405.03929 [cs].

[13] Y. Ren, X. Li, and W. Zhang. A Data-Driven Deep Learning Model for Weekly Sea Ice Concentration Prediction of the Pan-Arctic During the Melting Season. *IEEE Transactions*

on Geoscience and Remote Sensing, 60:1–19, 2022. Conference Name: IEEE Transactions on Geoscience and Remote Sensing.

[14] S. E. Stammerjohn, D. G. Martinson, R. C. Smith, X. Yuan, and D. Rind. Trends in antarctic annual sea ice retreat and advance and their relation to el niño–southern oscillation and southern annular mode variability. *Journal of Geophysical Research: Oceans*, 113(C3), 2008.

[15] L. Uebbing, H. L. Joakimsen, L. T. Luppino, I. Martinsen, A. McDonald, K. K. Wickstrøm, S. Lefevre, A.-B. Salberg, and J. S. Hosking. Investigating the Impact of Feature Reduction for Deep Learning- based Seasonal Sea Ice Forecasting.

[16] Y. Wang, X. Yuan, Y. Ren, M. Bushuk, Q. Shu, C. Li, and X. Li. Subseasonal Prediction of Regional Antarctic Sea Ice by a Deep Learning Model. *Geophysical Research Letters*, 50(17):e2023GL104347, 2023. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2023GL104347.

[17] J. Xu, S. Tu, W. Yang, S. Li, K. Liu, Y. Luo, L. Ma, B. Fei, and L. Bai. Icediff: High resolution and high-quality sea ice forecasting with generative diffusion prior, 2024.

[18] Z. Yang, J. Liu, M. Song, Y. Hu, Q. Yang, and K. Fan. Extended seasonal prediction of Antarctic sea ice using ANTSIC-UNet. *EGUsphere*, pages 1–25, June 2024. Publisher: Copernicus GmbH.

[19] Y. Zhu, M. Qin, P. Dai, S. Wu, Z. Fu, Z. Chen, L. Zhang, Y. Wang, and Z. Du. Deep Learning-Based Seasonal Forecast of Sea Ice Considering Atmospheric Conditions. *Journal of Geophysical Research: Atmospheres*, 128(24):e2023JD039521, 2023. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2023JD039521.