We would like to thank the reviewer for the constructive comments. Please find below our responses to the comments.

Review of "Calibration of short-term sea ice concentration forecast using deep learning"

Overview:

This study by Cyril Palerme and co-authors develops a U-Net deep learning to post-process sea ice concentration forecasts in the Arctic. The model uses predictors from numerical sea ice forecasts, weather forecasts, and satellite sea ice concentration observations, and their sensitivities are examined. Analysis of forecasts over the independent test period of 2022 indicate the deep learning model can outperform several noteworthy benchmark forecasts.

General comments:

I really enjoyed reading the paper and was encouraged by the results. The daily SIC forecast problem at <10 day lead time is a challenging one, and even state of the art numerical prediction models have a difficult time with it. So, it's encouraging to see how deep learning may be able to help in this regard.

Semi-major comments:

I have two semi-major comments, one of which might not be possible to address, and the other might not be a problem at all and just my ignorance. The first has to do with the verifying observation choice of the "new" AMSR2 product. I appreciate the analysis done in section 2.1 that compares this product against Norwegian ice charts, and I don't doubt that this is a fine product to use for the kind of the high resolution forecasts being produced. However, at such short lead times, I worry that there is an independence problem between one of the most important predictors in the U-Net model (AMSR2 SIC on the day preceding the forecast start date) and the verifying observations. Recall that the goal of the prediction problem is to predict the ground-truth state, not observed SIC, since observations contain random (and probably for SIC systematic) errors. While the systematic errors are much more difficult to address, it should be possible account for random errors in theory by using a different observational product for verification than was used as the predictors in the U-Net model. I realize this might be difficult given the restrictions on resolution for other products, and wanting to use an accurate product, but I would like to see the authors address this in the paper, ideally by using independent obs, but at the minimum raise it as a limitation of the study.

We agree with this comment. Our choice of comparing with the ice chart data was motivated by the fact that SAR data is the primary source for the ice chart analysis. In the prioritized list of information sources for ice charts, AMSR2 data ranks as 'h' in a list that runs from 'a' to 'i' (p. 34

in the JCOMM report), so the ice charts are nearly independent of AMSR2 data. The final paragraph in Section 2.1 has been rewritten to the following:

In addition, the ice charts produced by the Ice Service of the Norwegian Meteorological Institute (https://www.cryo.met.no/en/latest-ice-charts; JCOMM Expert Team on sea ice, 2017) are used as an independent dataset for evaluating the AMSR2 SIC observations and the forecasts developed in this study. The ice charts are manually drawn by ice analysts using several types of remote sensing data. Due to their high spatial resolution, synthetic-aperture radar (SAR) images constitute the main source of information where they are available. Elsewhere, visible and infrared observations are used in priority, while passive microwave retrievals are used where no other observations are available. For evaluating the SIC forecasts, the ice charts were interpolated on the grid used for the deep learning models using nearest neighbor interpolation. It is worth noting that the ice charts provide SIC categories and are not produced during weekends. Therefore, the number of ice charts available in 2022 for evaluating the SIC forecasts varies depending on lead time (between 144 and 243), and is considerably lower than the number of AMSR2 SIC observations available.

We added the figure below in the paper. This figure represents an evaluation of the ice edge position in the European Arctic using the ice charts from the Norwegian Meteorological Institute as reference.

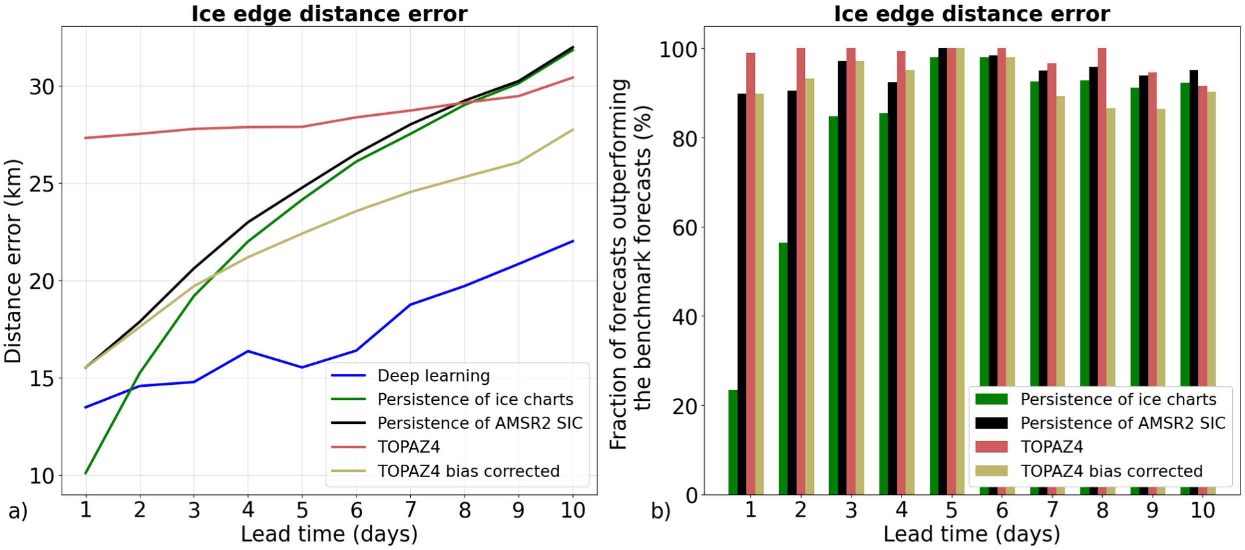

Figure 6. Performances of the deep learning models with the Attention Residual U-Net architecture during 2022 (test period) using the ice charts as reference. The ice edge position (defined by the 10 % SIC contour) is evaluated. a) Mean ice edge distance errors depending on lead time. b) Fraction of days in 2022 during which the forecasts from the models with the Attention Residual U-Net architecture outperform the different benchmark forecasts when the forecasts are evaluated using the ice edge distance error. It is worth noting that this evaluation is performed over the area covered by the ice charts from the Norwegian Meteorological Institute (European Arctic), and that the number of forecasts evaluated varies depending on lead time because ice charts are not produced during weekends.

Furthermore, we have added the following paragraphs for describing this figure:

In the section “3.3 Performances of the deep learning models”:

In order to assess the performances of the SIC forecasts using independent observations, an additional evaluation was performed in the European Arctic using the ice charts from the Norwegian Meteorological Institute as reference (figure 6). Since the ice charts provide sea ice categories (and not SIC as a continuous variable), only the ice edge position is evaluated in figure 6. On average, the forecasts from the deep learning models have an ice edge distance error 40 % lower than TOPAZ4 forecasts, 23 % lower than TOPAZ4 bias corrected, 29 % lower than persistence of AMSR2 SIC, and 22 % lower than persistence of the ice charts. While the forecasts from the deep learning models outperform TOPAZ4, TOPAZ4 bias corrected, and persistence of AMSR2 SIC for all lead times, they have worse performances than persistence of the ice charts for 1-day lead time (the ice edge distance error is 33 % larger). Moreover, only 23 % of the forecasts from the deep learning models outperform persistence of the ice charts for 1-day lead time. Nevertheless, the forecasts from the deep learning models significantly outperform persistence of the ice charts for longer lead times (p-value from the Wilcoxon signed-rank test < 0.05).

In the section “Discussion and conclusion”:

Using the ice charts from the Norwegian Meteorological Institute as reference, the forecasts from the deep learning models outperform all benchmark forecasts for lead times longer than 1 day in the European Arctic, but are worse than persistence of the ice charts for 1-day lead time. Since the deep learning models are trained using AMSR2 SIC observations for the target variable, it cannot be expected that they perform better than the differences between the two observational products (figure 1). While using ice charts for training deep learning models has been recently proposed by Kvanum et al., 2024, this does not allow to predict the SIC as a continuous variable.

The second is in regard to the Wilcoxon signed-rank statistical test used throughout the study to test the significance of differences in the scores for various models. I’m not familiar with this test, but I was surprised to see that some of the differences were found to be significant, such as at the 1 and 3 day lead times in Fig. 3a,b (but others too). Is there really such little variation in the errors from forecast to forecast that such small differences can be significant, or is there a problem with the test? Maybe the test has problems with autocorrelation in the errors from one week to the next? An alternative option would be a block bootstrap test. Can the authors comment on this concern and are they confident in the results of the test?

Thanks for pointing that out. We realized that we made a couple of mistakes with the interpretation of the Wilcoxon signed-rank test in the preprint. We have modified the following statements:

These differences are statistically significant (p-value from the Wilcoxon signed-rank test < 0.05) for all lead times and metrics, except for the ice edge distance error for 10-day lead time.
by:
These differences are statistically significant (p-value from the Wilcoxon signed-rank test < 0.05) for all lead times and metrics, except for the ice edge distance error for 9-day lead time.

and

When comparing the models using all predictors to those developed without TOPAZ4 sea ice forecasts, the differences in RMSE are statistically significant for all lead times, except 10 days.
by:
When comparing the models using all predictors to those developed without TOPAZ4 sea ice forecasts, the differences in RMSE are statistically significant for all lead times, except 1 and 10 days.

Furthermore, the Wilcoxon signed-rank test is an alternative to the Student t-test when the errors are not normally distributed (which is the case for sea ice concentration). The reason why there can be some confusion is that the Wilcoxon signed-rank test does not measure the statistical significance between the mean errors, but between the distribution of the errors. Therefore, it is sometimes possible that the mean errors are relatively close and that the Wilcoxon signed-rank test indicates that the differences are significant. In order to clarify this, we have added the following sentence at the end of section “2.4 Verification scores”:

It is worth noting that the Wilcoxon signed-rank test assesses the statistical significance between the differences in the distribution of the errors (and not between the mean errors).

Specific minor comments:

Title, abstract, L21, and throughout; I’ve only ever seen the term “calibration” in statistical post-processing of weather forecasts in the context of probabilistic/ensemble forecasts, in which part of the procedure is an adjustment on the ensemble spread or the shape of the forecast probability distribution. However, I’ve never seen it for deterministic forecasts like the ones under consideration here. The regression-based approaches to post-process deterministic weather forecasts are known as “model output statistics”, but there is no analogous term that I’m aware of yet for deep learning models. To avoid confusion with the probabilistic post-processing literature and methods therein, I think it would be more accurate to replace all instances of “calibration” with simply “post-processing”.

We agree with this comment and we have replaced “calibration” by “post-processing” in the paper. The title of the paper has also been changed, and the new title is “Improving short-term sea ice concentration forecasts using deep learning”.

L100; Can the authors be more specific when they say “the SIC trend calculated over the 5 days preceding the forecast start date”? What is meant by trend here?

We agree that this statement was not very clear. Therefore, we replaced the following statement:

“and the SIC trend calculated over the 5 days preceding the forecast start date.”

by:

“and the SIC trend calculated over the 5 days preceding the forecast start date (in % per day).”

L130; … “challenging areas” – again some specificity is needed here.

We agree with this comment and we have replaced the following sentence:

Furthermore, the impact of using attention blocks introduced by Oktay et al. (2018) in the decoder, and designed to improve predictions in challenging areas, is also evaluated.

by:

Furthermore, the impact of using attention blocks introduced by Oktay et al. (2018) in the decoder, and designed to give more weight (attention) on areas that are challenging to predict (these regions are identified by the attention blocks during training), is also evaluated.

L160; Just to say that I was glad to see this well thought out set of benchmark forecasts used.

Thank you !

Figure 2; The color bar is a bit misleading. Typically differentiating the range of SIC between 95% and 100% is not of any real practical interest, nor is the range between 0% and 15% (although noting the spurious values around 2% SIC in the text maybe noteworthy – typically values less than 15% are just clipped to 0%). Those small variations overwhelm the eye when looking at the maps and make the results look worse than they are. I suggest changing the increment to 5% across the full 0% to 100% range, as it would make any large differences between the maps more evident.

We agree with this comment and we have changed the color bar accordingly with increments of 5 % from 0 % to 100 %.

L268 and 269; I would avoid using the word “significant” when describing verification results unless one means “statistically significant”. It can be misleading.

We agree with this comment, and we have replaced the following sentences:

Figure 8 shows the seasonal variability in the performances of the deep learning models for lead times of 1, 5, and 10 days. Overall, the calibration shows robust results, with no significant seasonal cycle in the relative improvement compared to TOPAZ4 forecasts and Persistence. Moreover, the deep learning models significantly outperform all the benchmark forecasts for all the months, except in November when the 10-day forecasts are evaluated using the ice edge distance error.

By:

Figure 9 shows the seasonal variability in the performances of the deep learning models for lead times of 1, 5, and 10 days. Overall, the deep learning models show robust results, with no clear seasonal cycle in the relative improvement compared to TOPAZ4 forecasts and persistence of AMSR2 SIC. Moreover, the deep learning models outperform all the benchmark forecasts for all the months, except in November when the 10-day forecasts are evaluated using the ice edge distance error.

L276-277; Does this area of poorer performance in the East Siberian sea have a seasonal component to it (melt vs freeze)? It’s a good opportunity to bring up the fact that the use of only one year of test data makes it difficult to say if a feature like this is robust, especially if it’s only present in one of the seasons.

The forecasts from the deep learning models have better performances compared to TOPAZ4 in the melt season than in the freeze-up season in the East Siberian Sea. We have added the following sentence:

Nevertheless, it is difficult to determine if these poorer performances in the East Siberian sea are persistent because only one year is used for this analysis.