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.

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?

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".

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?

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

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

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.

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

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.