# Review of paper "Calibration of short-term sea-ice concentration forecasts using deep learning"

## 1 Content

The authors Palerme, Lavergne, Rusin, Melsom, Brajard, Kvanum, Sørensen, Bertino and Müller present a study on sea-ice concentration forecasts with lead times up to 10 days. They develop post-processing methods based on supervised machine learning and show that these outperform both TOPAZ4 and persistence sea-ice concentration forecasts. Also, they evaluate the impact of different types of predictors.

## 2 General comments

I thank the authors for submitting this already very mature manuscript, which I enjoyed reading and mostly had no problems comprehending. I got a bit lost when it came to the details of the model architecture, but that may be clearer to someone with a stronger AI background. If your target group is people with little AI background, I would suggest boiling down the technical description and spending a bit more time on explaining the terms (see detailed comments below). One aspect where I see room for improvement is the clarity of the motivation. Support of maritime operations is mentioned very prominently in both Abstract and Introduction. Yet, I find little information on crucial aspects for this like timeliness of the forecasts, infrastructure for providing the forecasts to ships operationally or the spatial resolution of the model outcome, compared to what is needed for navigational needs. So, I was wondering whether the current study is intended to be a proof of concept, with the applicability being the focus of future work. It would be good if the authors could clarify this, and elaborate upon for which other aspects their study is relevant besides supporting of maritime operations.

In summary, I suggest to accept this paper for publication after **minor revisions**.

## 3 Specific comments

**Abstract**

General impression: A very concise and yet comprehensive description of your paper. I find little to criticise.

L1: If operational support is your main motivation, I would suggest to add some brief information on the timeliness of your forecasts to the Abstract, as this is crucial for operational use. By how much does the post-processing delay the availability of forecasts, and does the accuracy increase achieved by post-processing outweigh this?

**Introduction**

General impression: A very good description of the existing literature on sea-ice forecasting. You could be a bit clearer on what the new aspects are that your study adds to the body of literature. I would also like to see a bit more details on the motivation and relevance of your study. It is certainly relevant, but it could be made clearer for what. Also, you could end the Introduction by formulating your research goals/questions, which can then be picked up in the Discussion and conclusion section.

L16: . . . is often limited by their inaccuracies: I agree, but a reference would be good nevertheless.

L17-19: If mentioned in Melsom et al. (2019), consider giving the accuracy for summer instead of a yearly mean, since that is when most of the maritime traffic happens. Large seasonal variability should still be mentioned.

L22: I find it a bit counter-intuitive to mention seasonal time scales when speaking about weather forecasts, which for me would be up to ten days. I would suggest to either simply speak about forecasts without specification on weather forecasts, or to refer only to those studies which focus on up to ten days lead time.

Last paragraph: You could be clearer on what precisely the goal of your study is. Which is the gap that you want to fill? Starting the paragraph with a sentence on what the new thing on your study is, and then elaborating on the why, would be a good transition towards the rest of the paper.

**Data**

General impression: Good to understand. Partly contained elements which I would suggest shifting to the Results or Methods section (see comments below).

L78: Good to see that you did this analysis, but I would move it to the Results section and focus on the data description here.

L95: Would be interesting to see later on if that impacts the forecast performance.

L97–116: Contains many elements where you describe what you do with the data, rather than which data you use. Consider shifting these elements to the Methods section, or merging Data and Methods section.

L97–105: I think it would help to add the categories in the table, and sort the table after categories, in the same order as they appear in the text.

L113: Does this influence the training, in the sense that the model is trained for points which are discarded later anyway, potentially diminishing the forecast quality for other,

"real" points?

## Methods

General impression: Section 3.1 was quite hard to comprehend due to the use of many technical terms, which I myself am not very familiar with. May be different for people with a strong AI background. Sections 3.2 and 3.3 were much clearer to me and do not need improvement, except for one question which I raise below.

L124–143: I find the description quite technical. Explaining, or, if they are not crucially needed, omitting terms like "residual connections", "convolutional/attention blocks" or "average pooling" would help to understand non-AI experts to follow your methodology.

L146: Does this also contain the "fake ocean points" described in L113, and if yes, does that influence the results?

L167: TOPAZ should be TOPAZ4 (also in other places, occurred several times).

## Results

General impression: As for the other sections, I find little to criticise. I did have some open questions, but nothing major.

L175: Does the varying number of model parameters influence the results, in the sense that you get more precise results for the Attention Residual U-Net models simply because there is a larger number of model parameters to optimize upon?

L184: Did you also see the spurious SIC which the Attention U-Net shows on other dates? Generally, concerning Fig. 2: It is fine to show only one day's maps as example, but it would be good to know if you also looked at other dates and how representative this date is for the overall development, how the RMSE's etc evolve over time. Since the ice refreezes rapidly at this time and the SIC is likely varying less in mid-winter, it would be interesting to see if you can also see that in the temporal development of the RMSE's and the comparison to your benchmarks.

L200–206: Do you know why the TOPAZ4 RMSE is higher than the persistence benchmarks for all lead times? Wouldn't you expect it to be better? It's not the focus of your paper, so don't spend too much room on this, but it was one of the first things which came to my mind when looking at Fig. 4.

L206ff: It would be interesting to see how removing predictors influences the runtime of the model. You could discuss, for example, whether the advantage gained in runtime outweighs the small benefit of including TOPAZ4 predictors, and if this might be reason enough to drop the TOPAZ4 predictors altogether.

Figure 5: It would be good if the colors of the bars would be consistent with those in Figure 4.

## Discussion and conclusion

The section provides a good summary of your findings, and puts them nicely into perspective with other studies. You could state your conclusions more clearly. This could for example be done by formulating a clear goal or research question(s) at the end of the Introduction which you can pick up and answer here to provide a nice framework for your paper. Also, I am missing a discussion on the practical applicability of your study. In the Abstract and Introduction, you name increasing marine traffic as one reason why short-term sea-ice forecasting is relevant. It would be interesting to discuss in how far your model is ready to support decision making for marine operations: How fast would your products be available? Do you have the means to transfer them in near-real time? Does the precision of your results meet the requirements of the onboard ship personnel? Or is your paper rather a proof of concept that it is possible to achieve high-quality short-term forecasts using deep learning, and the transfer to near-real time applications is something which could be done in future?