Response to first referee comments

We thank the reviewer for their useful comments on our manuscript. Our answers to the comments and modifications to the manuscript are provided below.

The manuscript addresses the critical need for accurate sea ice forecasting in the Arctic, driven by the increasing maritime activity due to sea ice retreat. A deep learning approach is developed that leverages operational atmospheric forecasts, ice charts, and satellite data to enhance short-term sea ice concentration forecasts within a 1 to 3 days timeframe, aiming for a detailed 1km resolution. The model's performance, validated against various thresholds of sea ice concentration contours, outperforms both baseline forecasts and two state-of-the-art dynamical sea ice forecasting systems across all considered lead times and seasons.

Nonetheless, the paper could stand to delve deeper into the model's limitations. Addressing potential biases from the training data and the effects of missing or inaccurate data could enrich the study. Suggestions for improvement are listed as below.

1. Place Table 1 within the 'Data' section for better context.

The table-positioning parameters have been updated to ensure that Table 1 is placed within the 'Data' section.

2. On page 6, line 140, provide clarification regarding the significance of the 'timeliness of 2.5 hours' for the AROME Arctic model, a detail omitted in Section 2.2.

We have modified the following sentence in Section 3.1:

In addition, AROME Arctic has a production time of about 2.5 hours, which ensures that the forecast initiated at 18:00 UTC are available before midnight, allowing us to publish deep learning forecasts on the same day as the input ice chart is published.

3. Using operational atmospheric forecasts, ice charts, and Sea Ice Concentration (SIC) from passive microwave observations as predictors is innovative. However, the paper should consolidate potential biases in these data sources and their impact on model performance in the discussion, making the article more logical and complete.

In the manuscript, when describing AROME Arctic we make sure to address that the system is operational and thus routinely receive updates which impacts distributional properties of predicted variables without retroactive effects. We also further described our choice of limiting training data to 2019 and onward as a direct response to avoid training on channels with differently distributed data. We have modified the following sentence in Section 2.2:

AROME Arctic has been in operation and continuous development since October 2015, routinely receiving updates which introduces permanent bias changes for predicted variables. Due to a major change to the representation of snow over sea-ice in 2018, a warm bias in near-surface temperatures above sea-ice was significantly reduced in the model (Batrak and Müller, 2019). Thus we start our training dataset at 2019 to avoid supplying our deep learning model with samples containing different temperature biases, especially close to the marginal ice zone (MIZ) where the greatest model response to predictors occurs.

Although Norwegian ice charts have little documentation regarding uncertainty estimation, we considered the comparison against AMSR2 as an analysis of the sensitivity to the sea ice product used for the target. Figure 5 shows that ice charts and AMSR2 have different occurrence frequency for different thresholds, and we show in our manuscript that initial differences between sea ice products are inherited by our deep learning system. We have modified Section 5 with the following to highlight this result:

However the analysis also suggests that inherent differences between sea ice products is reflected by deep learning forecasts, and we can not expect the forecasts to improve beyond that initial difference as the models are trained to only minimize the statistical error of their target sea ice product.

Yet, we disagree that addressing biases will strengthen our analysis. Since deep learning models learn to minimize the output error based on its input, as long as the data is consistently distributed over time, any biases would not impact performance since the model learn those as well. If distributional properties significantly changes in the training data, samples may contribute negatively or be neglected during training overall reducing the skill of the trained network. However as long as the data has a close to constant bias, all samples will contribute positively to the training as the relationship between output and the bias is part of what the model is being taught. We have modified the discussion (Section 5) to address the need to validate deep learning model performance for longer periods of time, since we believe that understanding how updates to physical models supplying predictors to a deep learning system impacts performance is crucial when considering operationalizing machine learning models.

Hence we recommend evaluating the forecasts with longer time series when they become available. With respect to the development of the operational weather prediction system AROME Arctic, a continued forecast evaluation can also facilitate understanding model

response to continuously updated atmospheric predictors and the potential of fine-tuning deep learning models.

4. On page 9, line 205, explain the rationale behind the selection of a specific number of epochs for model training.

We have added the following line to Section 3.3 in the manuscript:

We chose to train for 25 epochs as the validation loss rarely improved beyond that point.

5. The impact of hyperparameter tuning on model performance should be discussed. Were any automated hyperparameter optimization techniques like grid search or Bayesian optimization used?

We have modified section 4.1 Training performance and data considerations with a specification of what hyperparameters our grid search analysis was performed across.

The optimal U–Net width of 256 channels in the bottleneck was determined by performing a grid search on the validation dataset across learning–rate (0.0001 - 0.01) and U-Net depth (256 - 1024) (see Figure S2 in the supplement). To achieve consistent architectures between the developed models, we considered only variations of the 2-day target lead time model for the grid search and reused the results for models targeting all lead times.

We have also added the results from the grid search (Fig. R1) to the Supplement.

6. In section 4.2, the comparison with dynamical models should include a discussion on the computational efficiency of the deep learning model. This is particularly important for operational forecasting, where timely predictions are crucial.

We agree with the reviewer, and have added the following to section 4.2 addressing production time of Barents-2.5 in comparison with the deep learning forecasts:

Comparatively, a single member of Barents completes a 24-hour forecast in ≈1min, resulting in a 90% speed up when running on comparable hardware.

7. It would be beneficial to conduct a more detailed analysis of the model's performance across various sea ice concentration ranges in Section 4.2.

We agree with this comment, and have modified Figure 6 to present the Mean annual forecast error across different concentration thresholds (10, 40, 70 and 90%), similar to Figure 8 and Figure 9. The description of Figure 6 in Section 4.2 of the manuscript has also been updated to reflect this change:

We initially compare the deep learning forecasts against the baseline and dynamical forecasts in 2022 across all target lead times where we consider the yearly mean of the nIIEE for

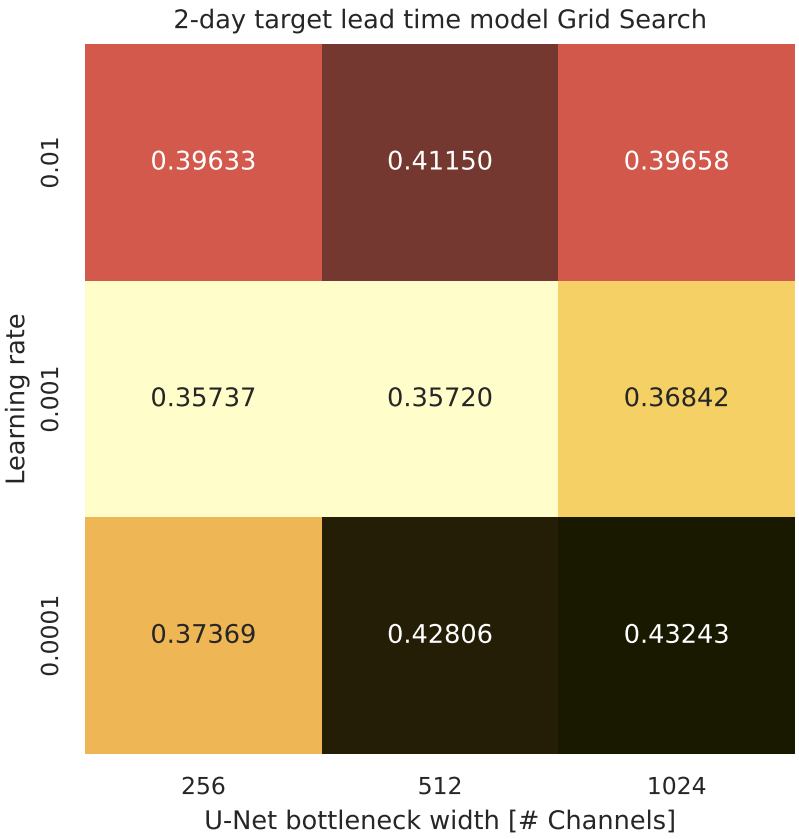

Figure R1: Grid search across varying learning rates and bottleneck widths for a deep learning model targeting 2-day lead time. The scores represent the minimum validational loss achieved before terminating training at 25 epochs.

different sea ice edge contours defined by (10, 40, 70 and 90%) concentration thresholds in Fig R2. For all considered lead times and concentration thresholds, the deep learning forecasts achieves the lowest nIIEE. Similar to persistence, nIIEE for the deep learning forecasts increases proportionally with lead time, although at a lower rate. Additionally, neither neXtSIM, free-drift nor the linear trend forecast are able to outperform persistence, on average for the 10% concentration contour scoring a factor of 1.57, 1.12, and 1.34 higher than persistence, respectively. Furthermore, the mean nIIEE between forecasts based on ice charts (Deep learning, Persistence and free-drift) and NeXtSIM and the linear trend whom are forced by a different sea ice concentration source is notably shifted from the 70% concentration thresholds and above. The nIIEE does not increase much with lead time especially for NeXtSIM when considering higher concentrations.

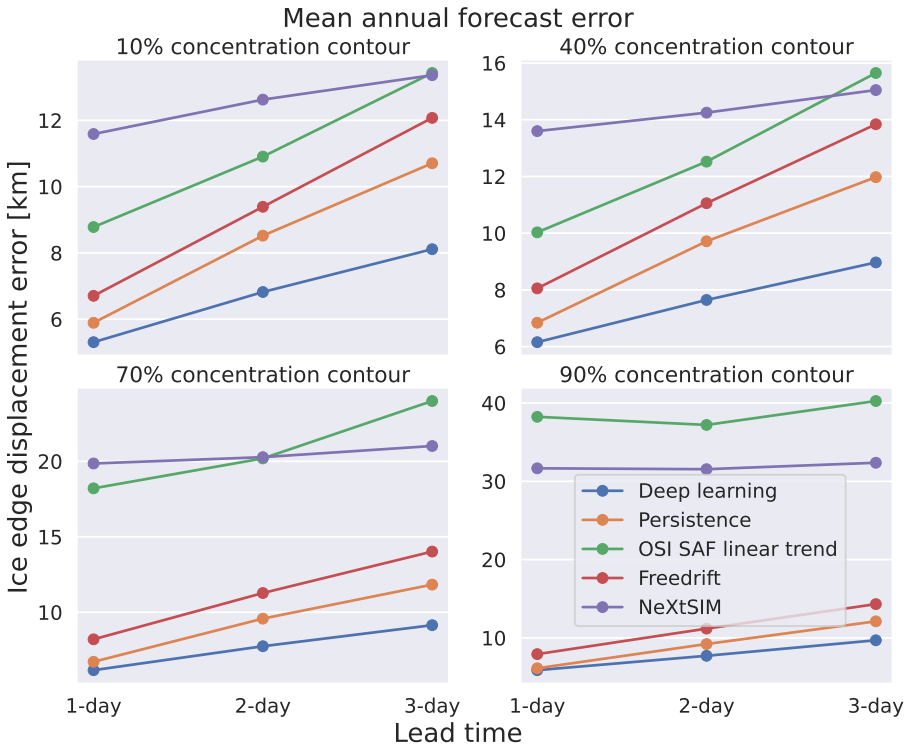

Figure R2: Mean annual ice edge displacement error as function of lead time for different sea ice concentration contours defined by 10, 40, 70 and 90% SIC. Only products with a complete coverage of 2022 have been considered.

The deep learning forecasts improve upon persistence by reducing the $nIIEE_{10\%}$ by a factor of 0.82. In terms of error-growth as a function of lead-time, the linear trend forecast is the only forecast where the slope of the error increases with increasing lead-time regardless of concentration threshold. This indicates that the linear trend from past OSI SAF SSMIS observations is unable to capture ice chart evolution especially for longer lead times. Moreover, the neXtSIM forecasts have the lowest error-growth with lead-time for lower concentrations, indicating that neXtSIM may provide more useful MIZ forecasts at longer lead-times.

8. Certain figures, especially those illustrating the model's performance compared to baseline and dynamical models, could be enhanced for clarity and aesthetics. For example, Figure 9 may require modifications to improve clarity.

We agree that Figure 9 is difficult to interpret and requires modifications to enhance its clarity. We have remade Figure 9 following the styles of Figure 6 and Figure 8, which preserves the content of the figure.

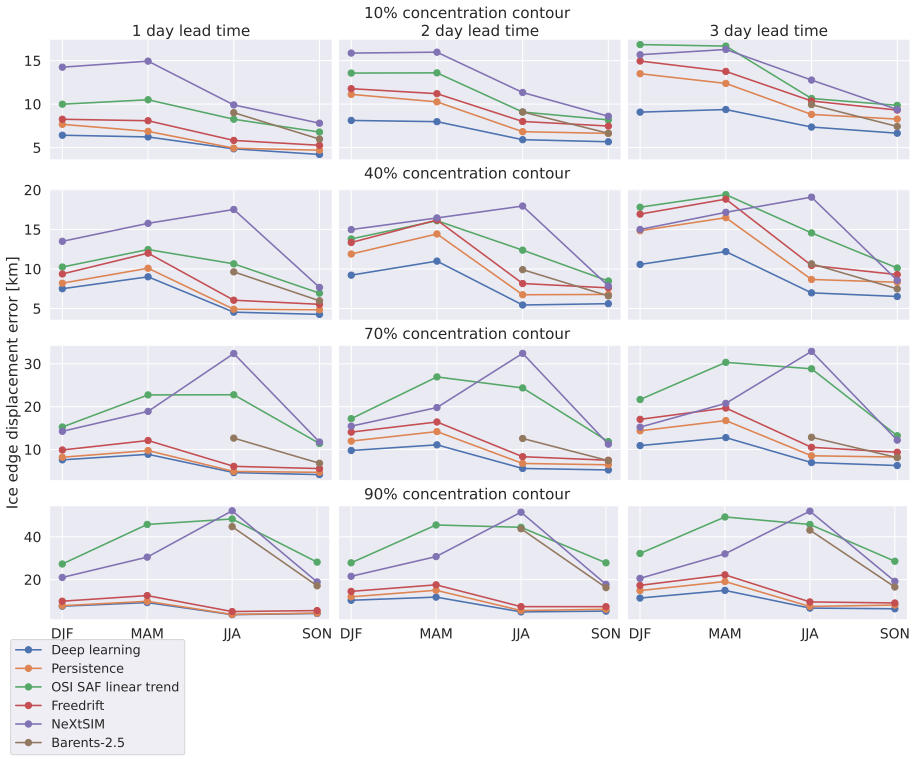

Figure R3: Model intercomparison for varying seasons, lead times and concentration contours. The ice charts are considered as reference. The values reported represent the integrated ice edge error normalized according to the length of the current SIC contour from the reference ice chart in km. The OSI SAF linear trend is computed from the past five days. Barents-2.5 results are only shown for summer and fall.