1 Model performance

The following section intends to explore the performance and capabilities of the deep learning system. Where the previous section 1.3 assessed the intra-training model performance, the current section will compare a benchmark deep learning model against baselines and physical models. The physical models have been previously described in section 1, and the baselines (although previously mentioned and to some extent utilized) will be derived in the following subsection. This section will first assess model performance against persistence. Afterwards, the deep learning system will be compared against other physical models. Setup and considerations will be described as they become relevant. Note that performance is commonly distributed seasonally, as in Winter (DJF), Spring (MAM), Summer (JJA) and Autumn (SON). Distributing the data seasonally was chosen to increase the robustness of each distribution, compared to a monthly distribution due to the limited number of samples for each month (Table 1).

1.1 Baseline-forecasts

Two types of baseline-forecasts are considered, persistence and a linear trend. The baseline-forecasts serve as a lower threshold which the Deep learning forecasts much outperform in order to be deemed skillful. A persistence forecast involves keeping the initial state of the system constant in time. Regardless of the forecast lead time the initial values for all grid cells are kept constant. Based on the analysis of autocorrelation if Section 1.2.1 and Figure 10, it is expected that persistence forecasts have some skill. For this work, a forecast has predictive skill if the forecast achieves a lower NIIEE than persistence, which is a similar approach as employed in Zampieri et al. (2019). We believe that using this threshold as the definition of a skillful forecast preserves the intent of validating the sea ice forecast in a manner relevant for maritime end users (Melsom et al., 2019; Veland et al., 2021).

The second baseline-forecast uses the linear trend, as described in section 1.2.3 and used as predictor for the deep learning system 1.1.3. However, the computed linear trend will be applied pixelwise to advance the initial state forward in time to a given lead time. As the linear trend is computed from OSI SAF SSMIS observations, it will consequently be applied to the same dataset. For clarity, the linear trend forecast is computed on the 1km AROME Arctic grid, and the computed values are clipped to match the valid value range, i.e. values $< 0 \rightarrow$ values $= 0 \wedge$ values $> 100 \rightarrow$ values $= 100$.

1.2 Verifying performance against persistence

For this section, a model with a depth of 256 channels in the final feature map, with a learning rate = 0.001 and all predictor variables have been used. Only the core training dataset was used for training, which include the years 2019 and 2020.

The seasonal distribution of NIIEE for all sea ice categories are shown for the deep learning system and persistence-forecast are displayed in Figure 1. For most of the sea-ice categories (>10%, >40%, >70% and >90%) the Deep learning system achieves a lower median, 25-th and 75-th percentile than the persistence-forecasts. For the (>0% and =100%) contours, the performance of the Deep learning system is inconsistent. In some seasons (Winter and Spring), the Deep learning system achieves lower median, 25-th and 75-th percentile compared to the (>0%) contour, however during Summer and Autumn Deep learning and persistence achieve similar NIIEE distributions. The Deep learning system is consistently outperformed by persistence-forecasts for the (=100%) contour.

Figures 2 and 3 shows the model confidence as an annual mean for all output contours (figure 2) and the (> 10%) contour distributed seasonally (figure 3). The confidence values shown are output pixel values after the sigmoid (equation 7), such that values closer than 1 are pixels that the model is more confident to belong in the output contour. Likewise, values closer to 0 are confident not to belong to the targeted contour.

Figure 2 shows that all cumulative contours, except = 100%, have a confidence pattern similar to the seasonal cycle of sea ice concentration. This is expected since for all contours and at all dates sea ice concentration is always present north of Svalbard and towards Greenland in the left of the model domain, whereas the less confident areas east of Svalbard exert spatial variability since the sea ice drifts and accumulates / melts. However, it is noted that the = 100% contour (Figure 2f) shows a lower overall confidence level, as well as only being restricted to the land structures present in the scene.

The seasonal confidence cycle for the = 100% contour is shown in figure 3. It is seen that the spatial distribution of confidence tends to resemble the land-covered pixels for all seasons, although with varying levels of confidence. E.g. Novaya Semlya is barely visible in Figure 3d.

The monthly mean sea ice edge length for the sea-ice charts and the predictions is shown in Figure 4. The predicted ice edge follows a similar seasonal pattern to the ice edge length from the target ice charts. Each monthly mean predicted sea ice edge length is biased towards shorter lengths, and the bias increases with longer forecast lead-times.

Moreover, the monthly distribution of the different sea ice categories is shown in Figure 5. The figure shows that the Deep learning model resolve the area of each contour with a similar scale and variability as the target sea ice charts.

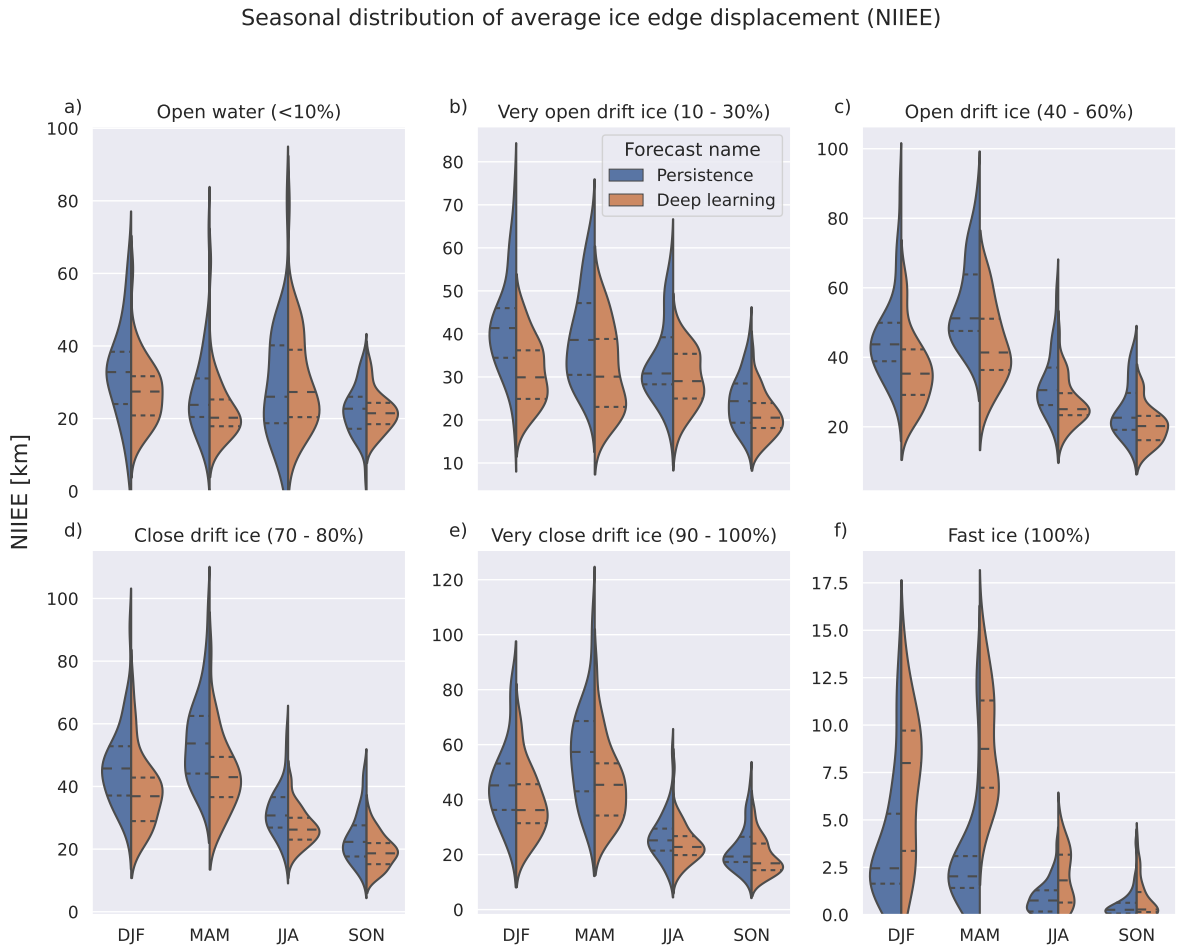

Figure 1: Seasonal distribution of the mean ice edge displacement NIIEE for the different sea ice chart categories. The sea ice concentration range for each contour which denote the lower concentration threshold is also noted. The lower and upper dashed line denote the interquartile range, with the middlemost dashed line showing the distribution median.

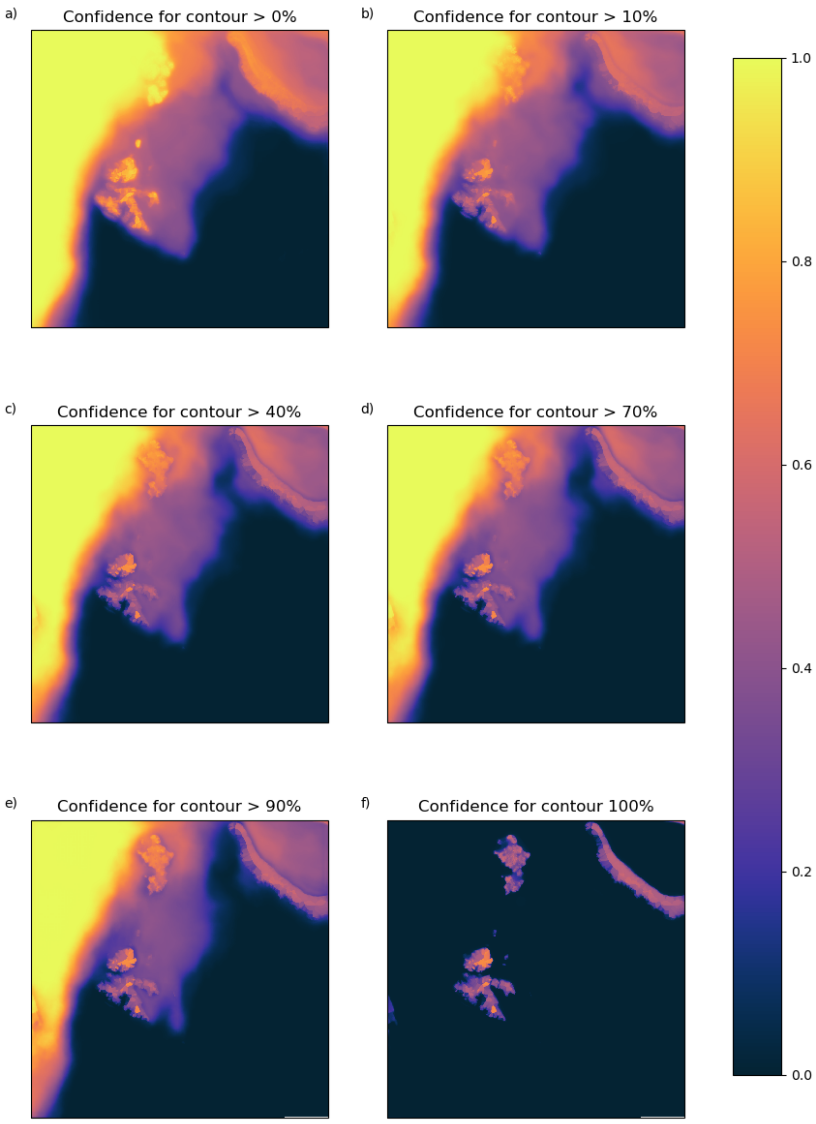

Figure 2: Mean annual probabilities for the different cumulative contours outputted by the model (the class ice free open water is not shown).

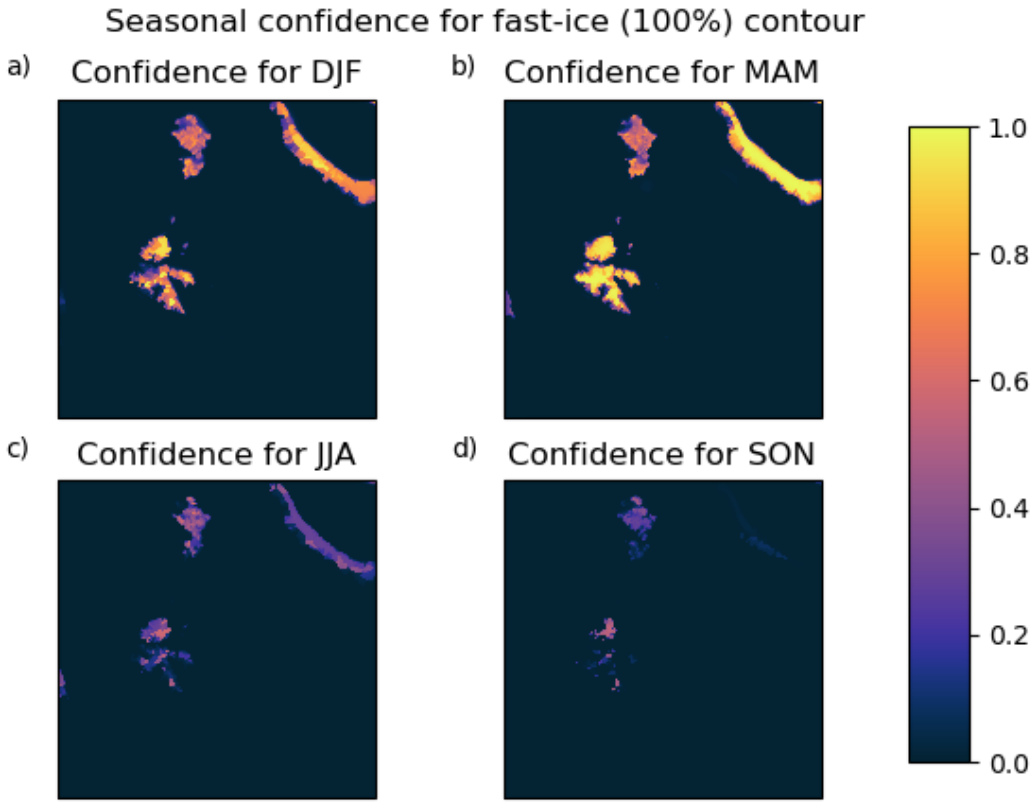

Figure 3: Mean seasonal confidence for the (= 100%) cumulative contour.

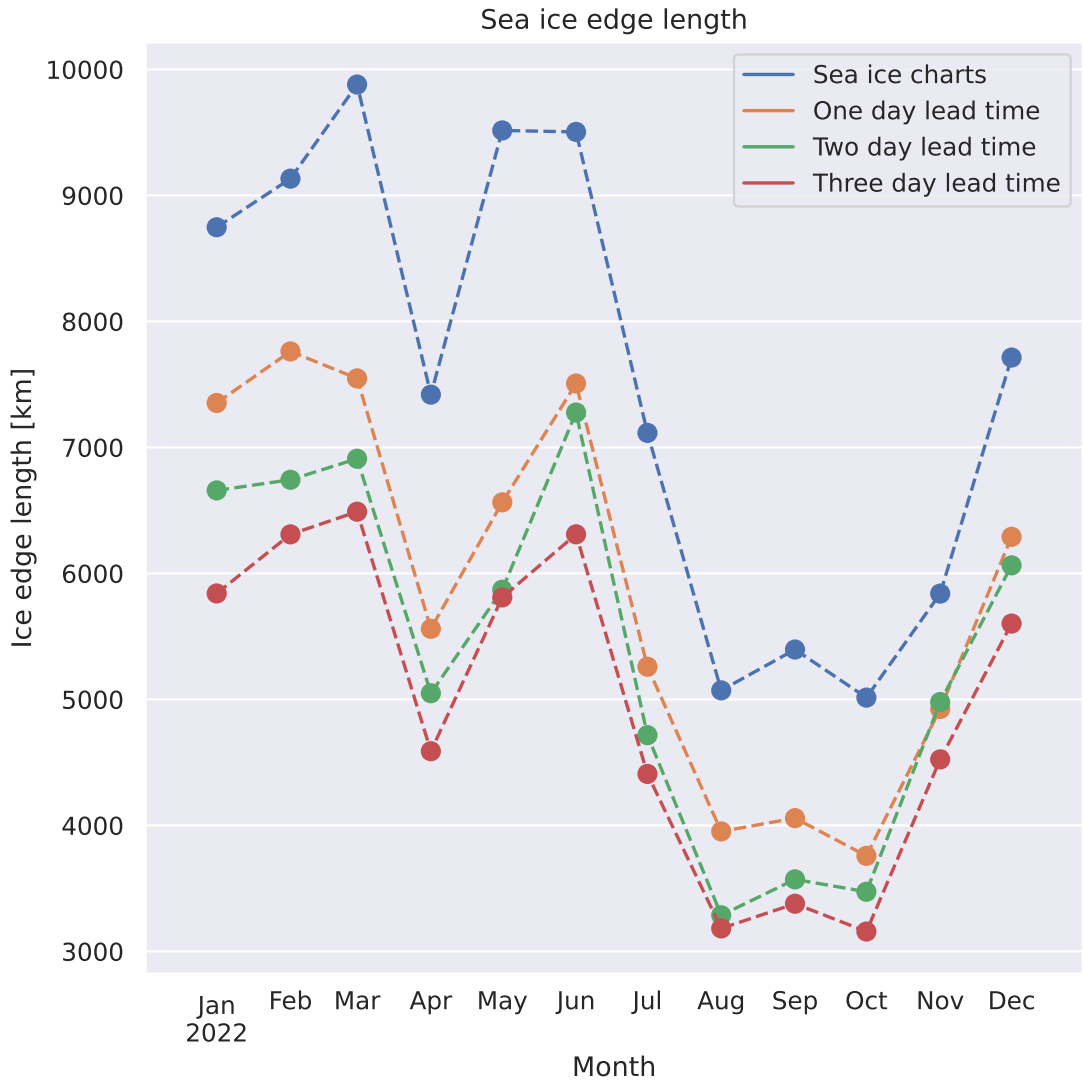

Figure 4: Mean monthly sea ice edge length for the entire 2022 test dataset. The ice edge is defined from a 10% threshold, which results in the 10% contour being used to define the ice edge. Each entry in the defined sea ice edge are on a 1km resolution. Each deep learning marker is annotated with the mean monthly bias with respect to the target sea ice edge length.

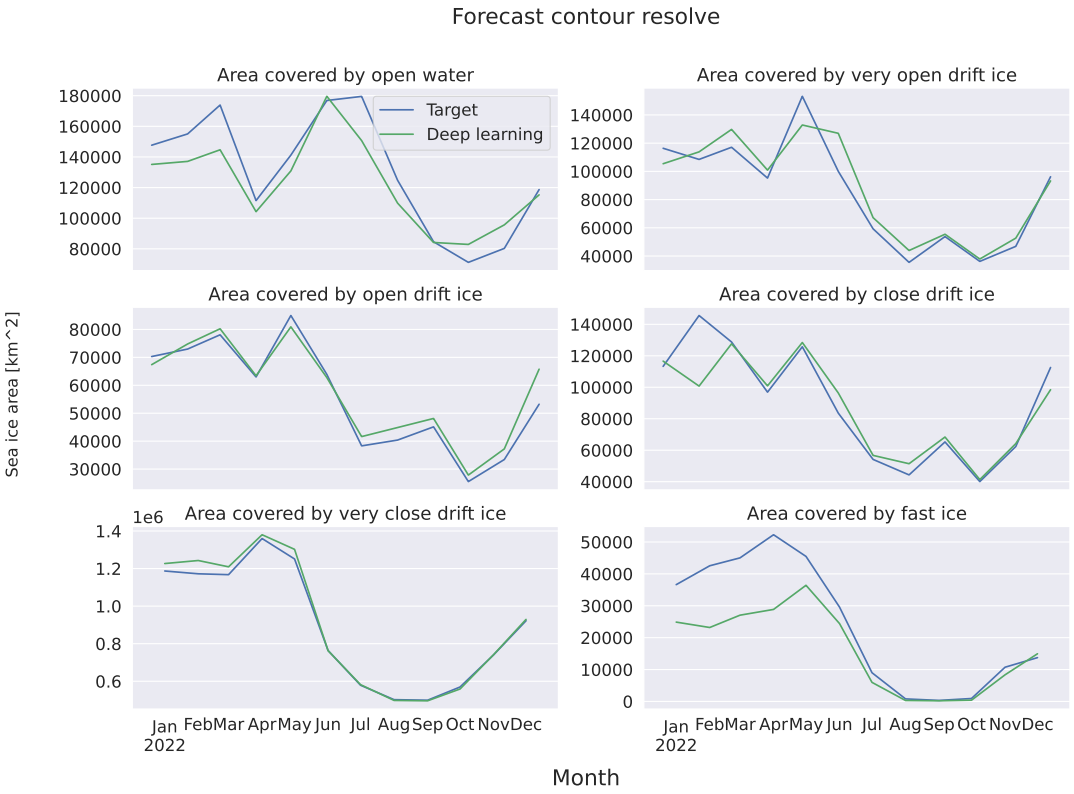

Figure 5: Mean monthly sea ice category distribution for the model and the target sea ice charts for the 2022 test dataset. Each contour is represented by the sea ice area, which is computed from the sum of pixels in each contour times their spatial extent.

1.3 Intercomparison of sea-ice forecasts

This section covers results regarding the multi-product comparison. First, the preparation of samples as well as setup of the comparison environment is described. The physical models considered for this comparison are neXtSIM (Williams et al., 2021) presented in section 1.3.2 and Barents-2.5 (Röhrs et al., 2023) presented in section 1.3.3, whereas the considered baseline-forecasts are persistence and the linear sea ice concentration trend described in section 1.1. Two different products are used as ground truth. The first product is the sea ice charts, which will be utilized similarly as when comparing only against persistence in section 1.2. The second product to be utilized as ground truth is the independent AMSR2 observations produced by Spreen et al. (2008).

When comparing against multiple products, the coarsest resolution model is used as a common spatial resolution. Also, the projection of the coarsest resolution is used for all products, such that other products have to be interpolated onto the grid of the coarsest resolution model, which is done using nearest neighbor interpolation. As both baselines-forecasts have a daily forecast frequency, comparing either with a deep learning prediction involves identifying the forecast with similar bulletin- and valid date, i.e. initialized at the same day and targeting the same lead time. When utilizing the sea ice charts as the ground truth, the spatial resolution of neXtSIM (3km) is the coarsest, and thus all products are interpolated onto the same resolution.

Comparing against the two physical models requires a consideration of the hourly forecast frequency (Williams et al., 2021; Röhrs et al., 2023) of both models. First, given a published sea ice chart, the comparable physical model is initialized the following day at 00:00 UTC. Furthermore, a daily mean is computed from the 24 steps forward in time taken by the physical model when it covers the valid date of the deep learning forecast. Even though the sea ice charts only convey information about the sea ice concentration up until their publication time, the operational product is considered a reference for the entirety of the publication date. Moreover, to reduce introducing a bias towards the time of day to the physical forecasts as well as limiting the spatial variability induced by the lack of a temporal mean, reducing the physical forecasts to daily averages is considered a more comparable approach than e.g. selecting a single hour (15:00 UTC) from the forecasts.

Since the AMSR2 observations are supplied on a 6.25 km spatial resolution (Spreen et al., 2008), when AMSR2 is used as the ground truth all data is interpolated to match the resolution of AMRS2. Although the AMSR2 data have a substantially coarser spatial resolution compared to the sea ice charts or the deep learning system, the data makes it possible to assess the generalizability of the deep learning performance when targeting an unseen and independent ground truth.

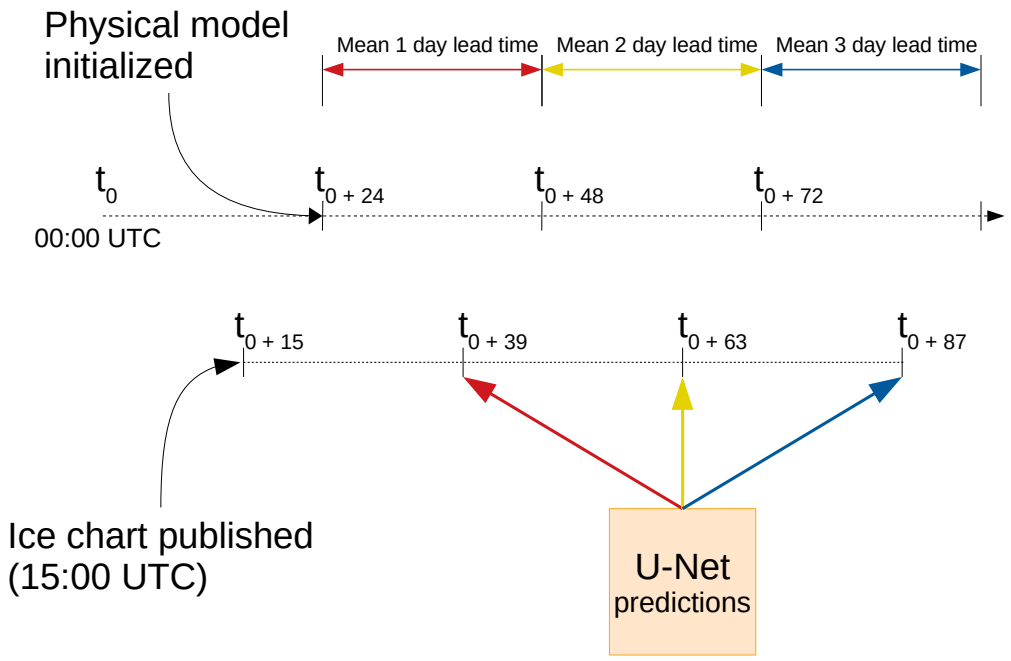

Figure 6: Overview describing how a physical model with an hourly frequency is compared against a deep learning forecast. Timestamps are hourly, and relative to 00:00 UTC the day a sea ice chart is published. The physical model is initialized the following day. Colors are used to denote lead time comparability, with red = 1, yellow = 2 and green = 3 day lead time.

From this setup, the mean of the first 24 hours of a forecast from a physical model is compared against a deep learning prediction with one day lead time, the mean between 24 and 48 hours are compared against a deep learning prediction with two day lead time and the mean of the third predicted day is compared against a deep learning prediction with three day lead time. Figure 6 summarizes the process. Note that Barents-2.5 only has a 66 hour lead time (Röhrs et al., 2023), thus the mean between $t = 48$ and $t = 66$ is computed when comparing against a three day lead time prediction.

It is noted that when comparing against multiple forecast products as described in figure 6, only the common dates shared between all products are used. With the current setup, where neXtSIM, Persistence, Deep learning, OSI SAF trend and Barents-2.5 are considered, the test dataset is reduced from 196 to 171 samples, 147 to 130 samples and 142

to 125 samples for 1, 2, and 3 day lead time respectively. Moreover, Barents-2.5 is only considered starting with the month of June, to comply with the spin up time of its data assimilation system (Röhrs et al., 2023).

Figure 7 shows the seasonal distribution of NIIEE for the different forecast systems and benchmarks, following the setup described in figure 6. By inspecting figure 7, it can be seen that only the products based on the sea ice chart are able to achieve consistently low NIIEE for the $> 0\%$ contour. Furthermore, for the $\geq (10, 40, 70, 90)\%$ contours, the deep learning system achieves the lowest median and mean values compared to all the other products. It can also be seen that neXtSIM tends to increase its mean and median as well as spread for increasing contours, with a similar although not as consistent pattern for Barents-2.5. Moreover, the OSI SAF trend typically has the highest valued outliers in the displayed ranges. Finally, no product is able to achieve a lower mean or median NIIEE compared to persistence when inspecting the 100% (fast ice) contour.

The fraction of days where the Deep learning system achieves lower NIIEE compared to each considered product is shown in Figure 8. The figure shows that the deep learning system consistently achieves a $\geq 50\%$ success rate compared to all products, except for persistence-forecast with 1 day lead time in July, August and September as well as Barents-2.5 2 day lead time in November and December. When compared to neXtSIM at 1 day lead time (figure 8 (a)), the Deep learning system achieves a lower NIIEE at all considered dates in the test data. However, it can also be seen that a lower amount of days with lower NIIEE than neXtSIM are achieved as the lead time increases. The same pattern may also be seen in the Barents-2.5 data as the mean fraction of days with lower NIIEE for the Deep learning system also decrease with lead time, although Barents-2.5 is only able to achieve lower NIIEE more than 50% of the dates for a 2 day lead time as previously noted. With respect to persistence, the Deep learning forecasts seem to achieve a higher fraction of days with lower NIIEE as lead time increases, although there is no trend for the individual months. At the ($\geq 10\%$) contour, the OSI SAF trend is consistently beat by the deep learning system during Winter and Spring, with less consistency observed during the Summer and Autumn seasons.

The spatial distribution of product error is shown in Figure 9. From the figure, it can be seen that both products which are based on the sea ice charts (Deep learning system and persistence-forecasts) have lower bias than the three other products, as well as only exerting biases in the MIZ. Moreover, it can be seen from the top row in figure 9 the neXtSIM data have a negative bias along the sea ice edge, which is prominent during Winter and Spring. Moreover, the OSI SAF trend seem to have a strong negative bias along a wide sea ice edge. Finally, Barents-2.5 seem to have a positive bias around Svalbard in the Summer, with a less prominent overall bias during the Fall.

The seasonal NIIEE distributions shown in figure 10 is created similarly as figure 7, but

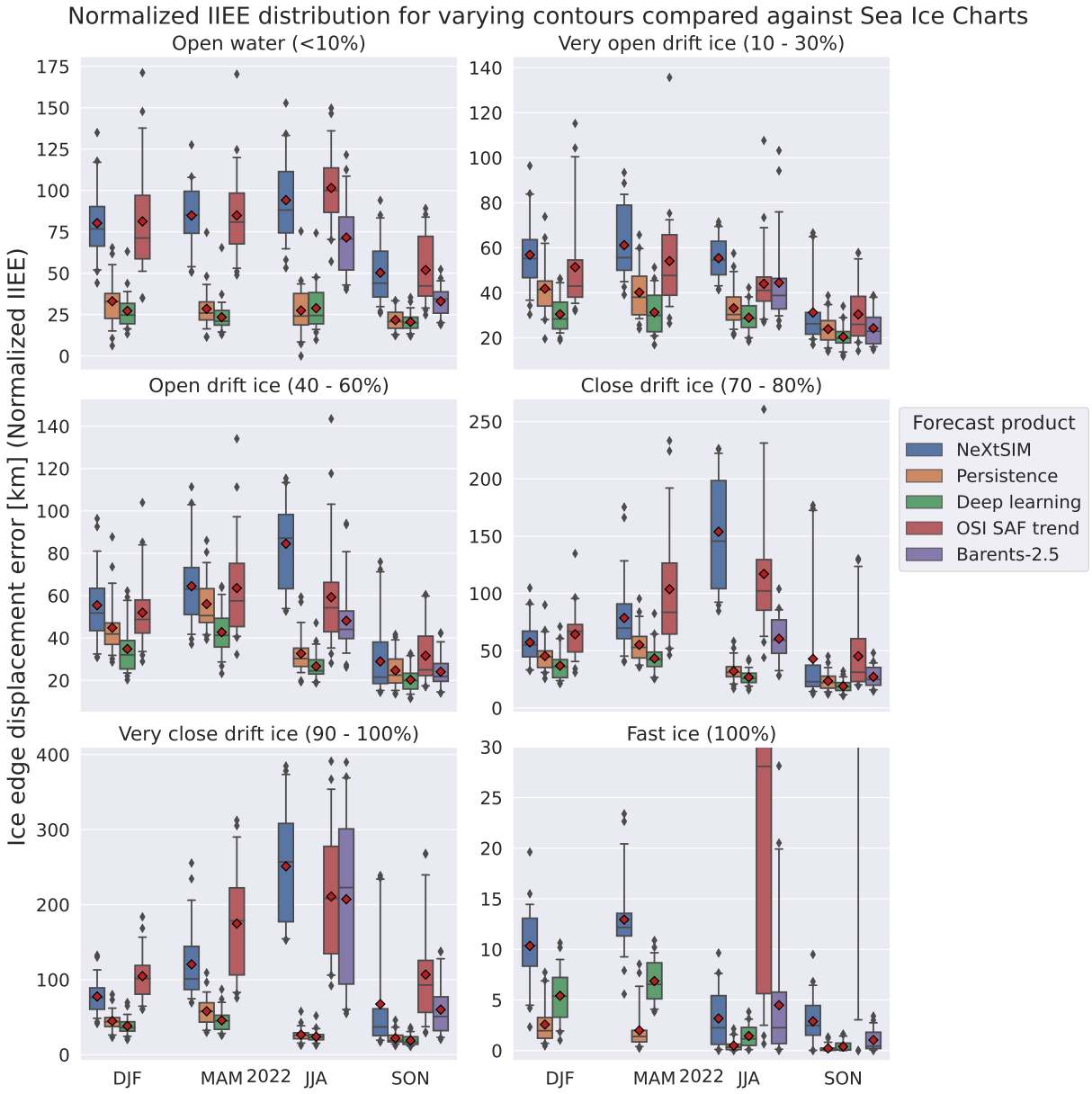

Figure 7: Model intercomparisson with a two day lead time. The boxes are constructed from seasonally distributed NIIEE values computed from the test dataset (2022). The sea ice charts are considered as targets. Each box cover the interquartile range (25th - 75th percentile), with whiskers covering the 5th and 95th percentile. The line in each box is the median, and the red diamond is the mean. The IIEE is normalized according to the climatological sea ice edge at the forecast valid date. The extent of the y axis is limited in such a way that the distributions are easily readable, at the expense of some outliers not being visible. The OSI SAF trend is computed from the past 7 days.

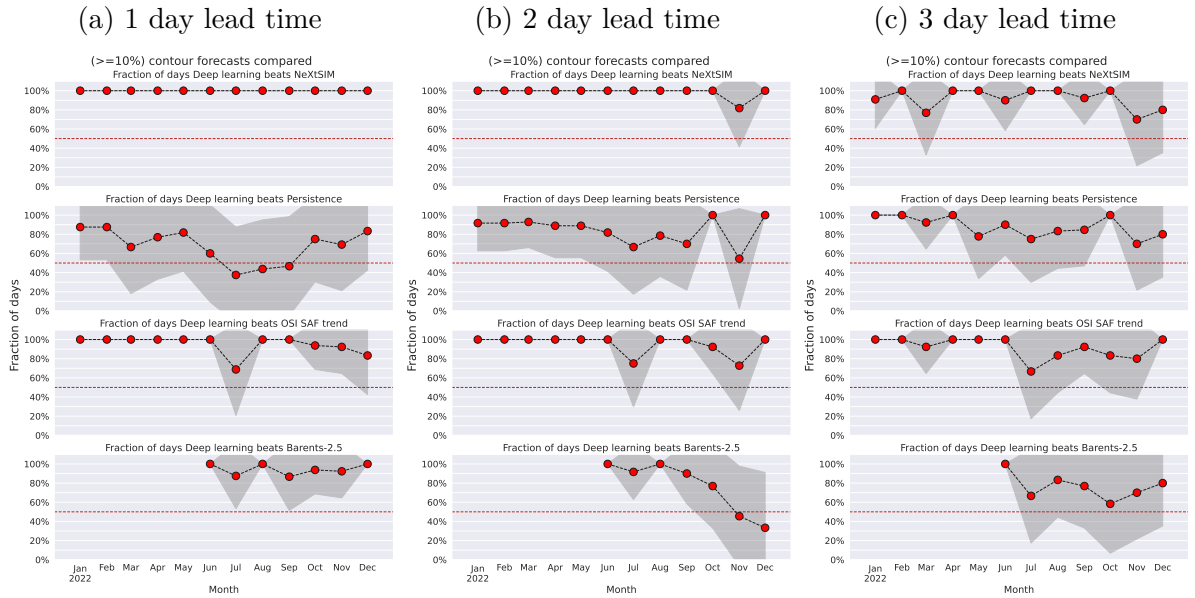

Figure 8: Fraction of days where the Deep learning forecast achieves a lower NIIEE than the compared product, distributed monthly for all lead times. Only the ($\geq 10\%$) contour has been considered, due to the relevance of the contour with respect to the definition of the sea ice edge and its application to operational end users. The red dashed line denotes the 50% line. Gray contours denote the uncertainty (standard deviation) for each month. The sea ice chart has been used as ground truth target when computing the IIEE, and the score has been normalized according to the climatological sea ice edge.

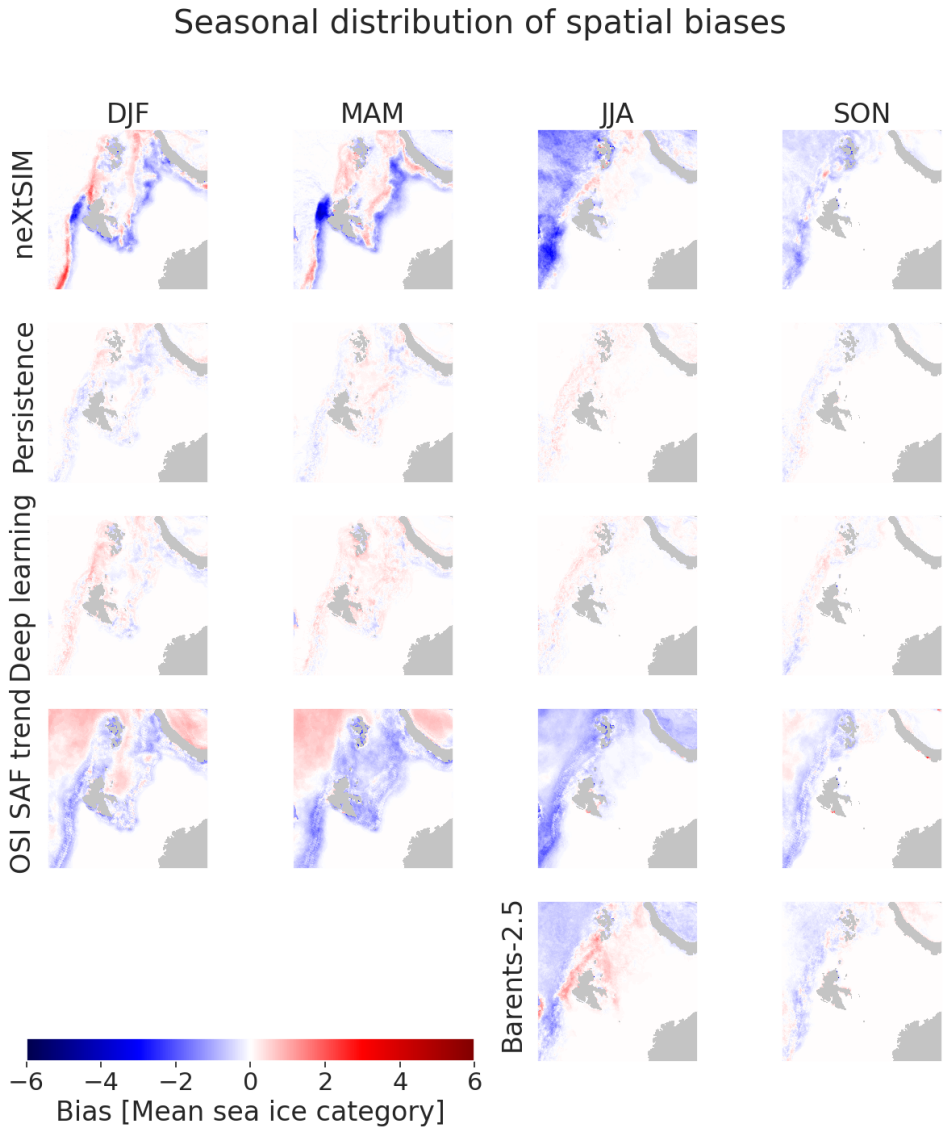

Figure 9: Spatial distribution of the mean seasonal error for predicted sea ice categories between the compared products. The data is interpolated onto the neXtSIM grid, and the test dataset is considered.

with AMSR2 as the ground truth data, which also implies that all data have been interpolated onto the 6.25km AMSR2 grid (Spreen et al., 2008). Contrary to what was observed in figure 7, the both the Deep learning system and persistence-forecasts in figure 10 exert significantly higher NIIEE at the $> 0\%$ contour. However, Barents-2.5 also exert a similar increased NIIEE as the Deep learning system and persistence-forecasts at the same contour. Moreover, both the Deep learning system and persistence-forecasts are within the interquartile range of neXtSIM and OSI SAF trend starting at the ($\geq 10\%$) contour. At the 0 and 10% contours, the OSI SAF trend exerts the lowest mean and median NIIEE for all months except SON where neXtSIM achieves the lowest median and mean. However, starting at the ($\geq 40\%$) the deep learning system has the lowest median and mean NIIEE, which lasts until the 100% contour where performance is comparable between all products except for the OSI SAF trend during Winter and Spring.

Following the result seen in the upper leftmost distribution in Figure 10, Figure 11 shows a comparable figure but with a deep learning model which does not predict the 10% and 100% contours as described in section 1.3.4. By inspecting the >0% contour, it can be seen that the deep learning system achieves significantly lower NIIEE than persistence, as well as the deep learning system in figure 10. Otherwise for the other contours, the performance of the deep learning system is comparable to the deep learning system in figure 10.

The boxplots in Figure 12 computes the >0% contour NIIEE against AMSR2 with the model used in Figure 10 but with the predicted >0% contour removed. The distribution seen in the figure resembles that in Figure 11, with the Deep learning forecasts performing significantly better than persistence.

Kanskje dette også skal i appendix?

References

Melsom, A., Palerme, C., and Müller, M.: Validation metrics for ice edge position forecasts, Ocean Science, 15, 615–630, https://doi.org/10.5194/os-15-615-2019, 2019.

Röhrs, J., Gusdal, Y., Rikardsen, E., Moro, M. D., Brændshøi, J., Kristensen, N. M., Fritzner, S., Wang, K., Sperrevik, A. K., Idžanović, M., Lavergne, T., Debernard, J., and Christensen, K. H.: Barents-2.5km v2.0: An operational data-assimilative coupled ocean and sea ice ensemble prediction model for the Barents Sea and Svalbard, Geoscientific Model Development, https://doi.org/10.5194/gmd-2023-20, 2023.

Spreen, G., Kaleschke, L., and Heygster, G.: Sea ice remote sensing using AMSR-E 89-GHz channels, Journal of Geophysical Research, 113, https://doi.org/10.1029/2005jc003384, 2008.

Veland, S., Wagner, P., Bailey, D., Everett, A., Goldstein, M., Hermann, R., Hjort-Larsen, T., Hovelsrud, G., Hughes, N., Kjøl, A., Li, X., Lynch, A., Müller, M., Olsen, J.,

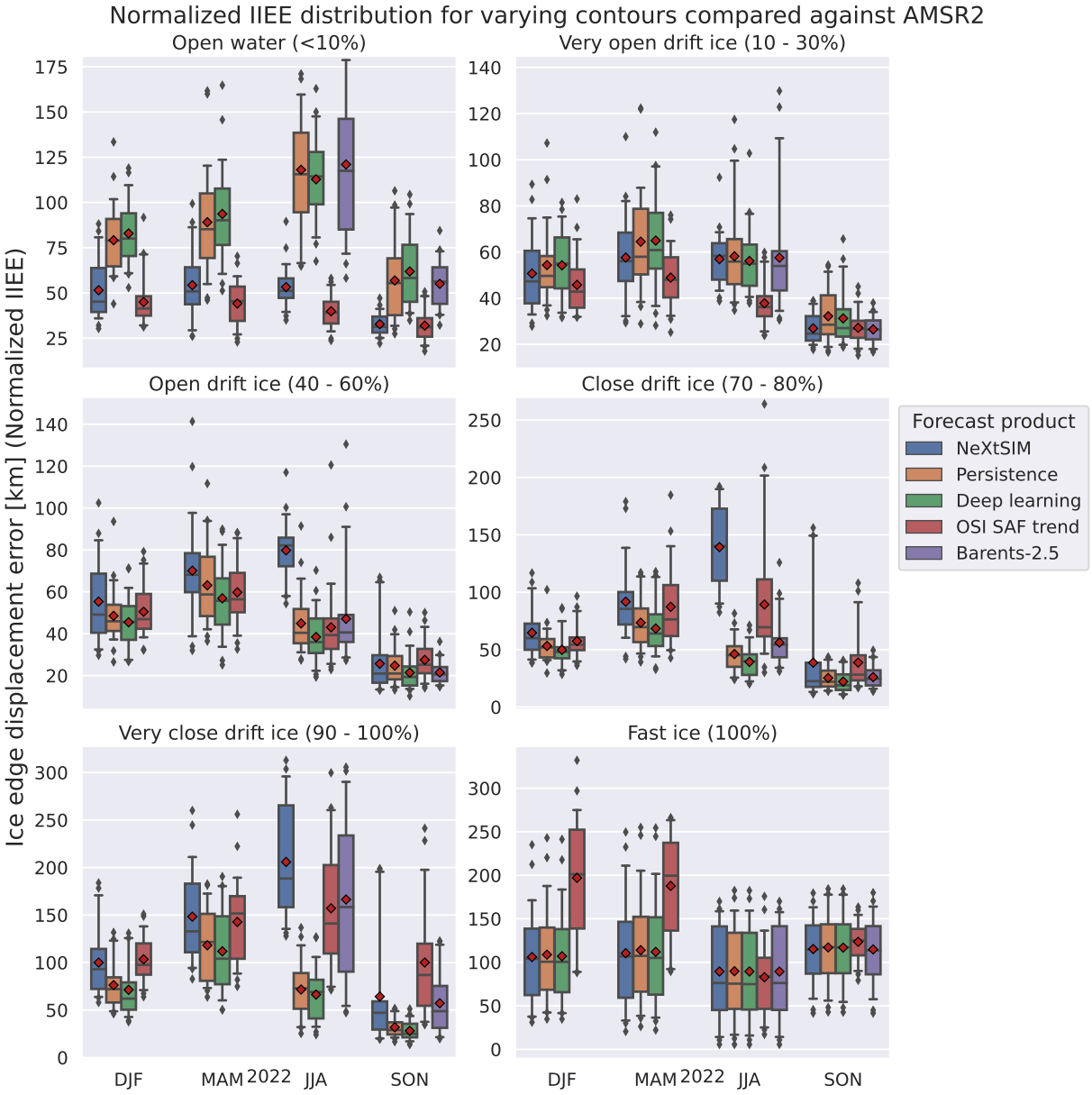

Figure 10: Same as figure 7, but with AMSR2 sea ice concentration as the ground truth data. Note that AMSR2 is only used as reference for validation, not as target variable for training the Deep learning system (Section 1.3)

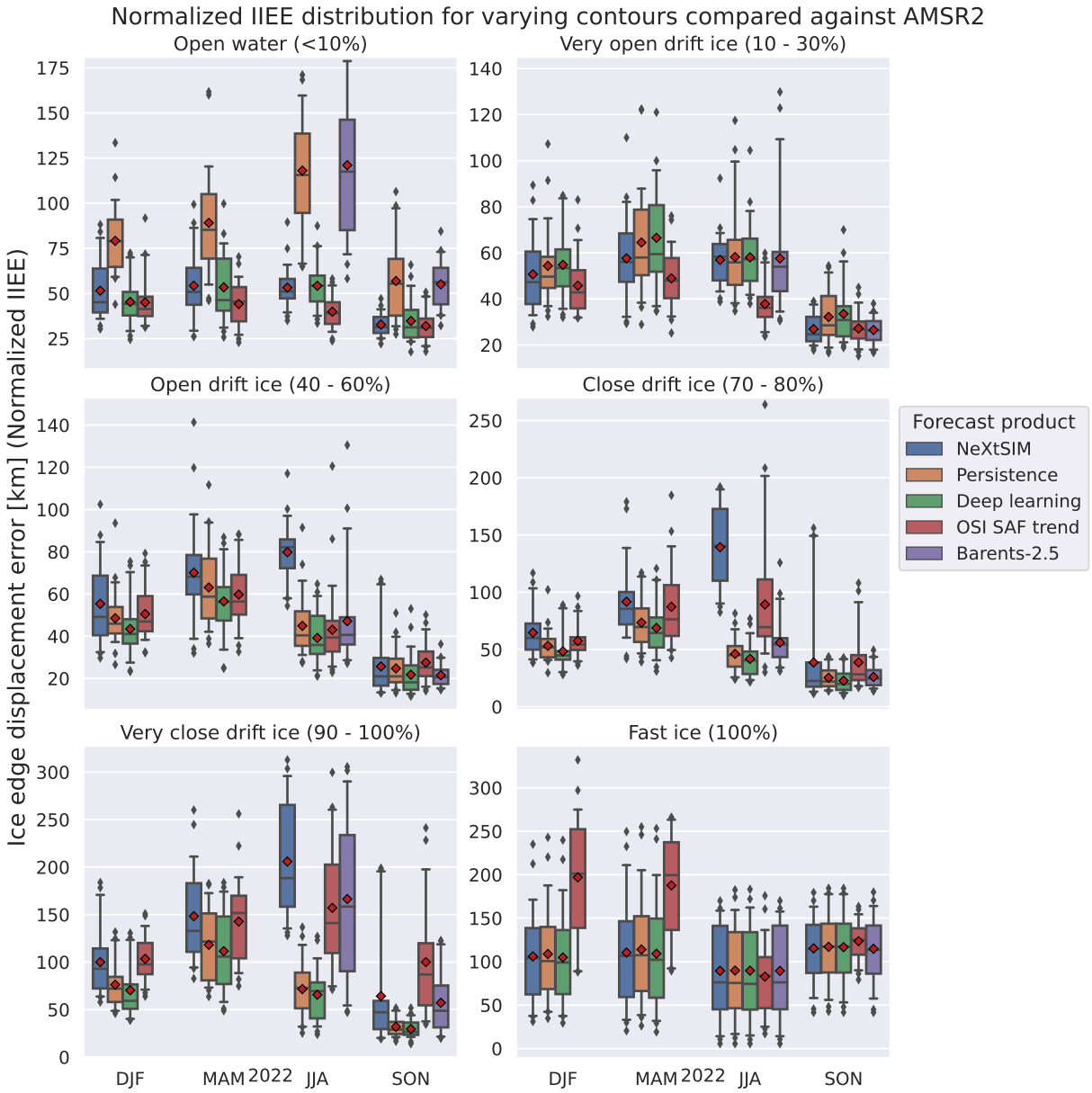

Figure 11: Same as figure 10, but the deep learning system used has reduced output classes.

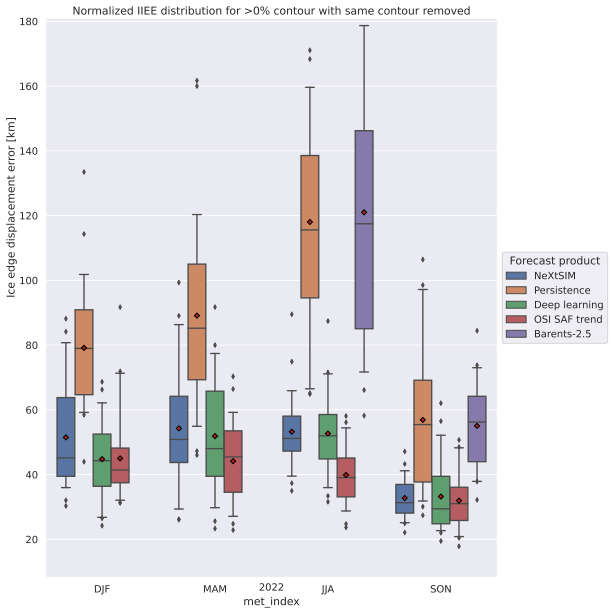

Figure 12: NIIEE for the >0% contour with the model from Figure 10, but with the values in the >0% contour set to category 0 (ice free open water)

Palerme, C., Pedersen, J. L., Rinaldo, ., Stephenson, S., and Storelvmo, T.: Knowledge needs in sea ice forecasting for navigation in Svalbard and the High Arctic, Tech. Rep. NF-rapport 4/2021, Svalbard Strategic Grant, Svalbard Science Forum, 2021.

Williams, T., Korosov, A., Rampal, P., and Ólason, E.: Presentation and evaluation of the Arctic sea ice forecasting system neXtSIM-F, The Cryosphere, 15, 3207–3227, https://doi.org/10.5194/tc-15-3207-2021, 2021.

Zampieri, L., Goessling, H. F., and Jung, T.: Predictability of Antarctic Sea Ice Edge on Subseasonal Time Scales, Geophysical Research Letters, 46, 9719–9727, https://doi.org/10.1029/2019gl084096, 2019.