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.

1.1 Baselines

To types of baselines are considered, persistence and a linear trend. A persistence forecast is constant in time. Regardless of the forecast lead time the initial values for all grid cells are kept constant. Moreover, the autocorrelation of sea ice concentration from the sea ice charts was shown in section 1.2.1 to be high for short lead times.

The second baseline 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. $\text{values} < 0 \rightarrow \text{values} = 0 \wedge \text{values} > 100 \rightarrow \text{values} = 100$.

I discussion nevne at persistence er vanskelig å slå, vise resultater fra (Zampieri et al., 2019)

1.2 Verifying performance against persistence

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

The seasonal distribution of average ice edge displacement for all sea ice categories found in the sea ice charts are shown for the deep learning system and persistence are displayed in figure 1. Figure 1 demonstrates the predictive performance for the deep learning system measured at each resolved contour. In figure 1 b), c), d) and e), the deep learning system achieves a lower median 25-th and 75-th percentile than persistence.

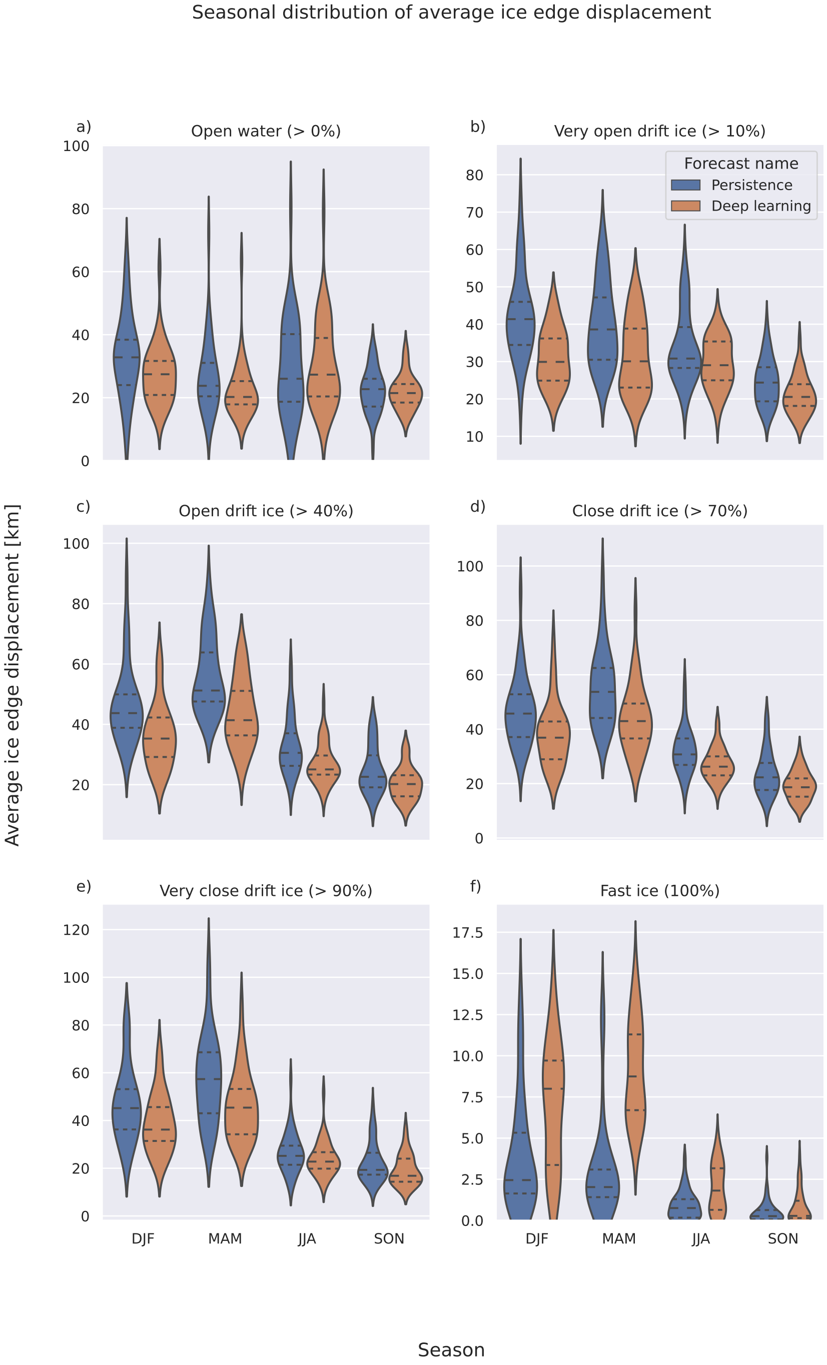


Figure 1: Seasonal distribution of the mean ice edge displacement (Normalized IIEE) for the different sea ice chart categories in the form of cumulative contours. The related sea ice concentration range for each contour is also included. The lower and upper dashed line denote the interquartile range, with the middlemost dashed line showing the distribution median.

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 outputted contour. Likewise, values closer to 0 are confident not to belong to the targeted contour.

1.3 Inter-product comparison

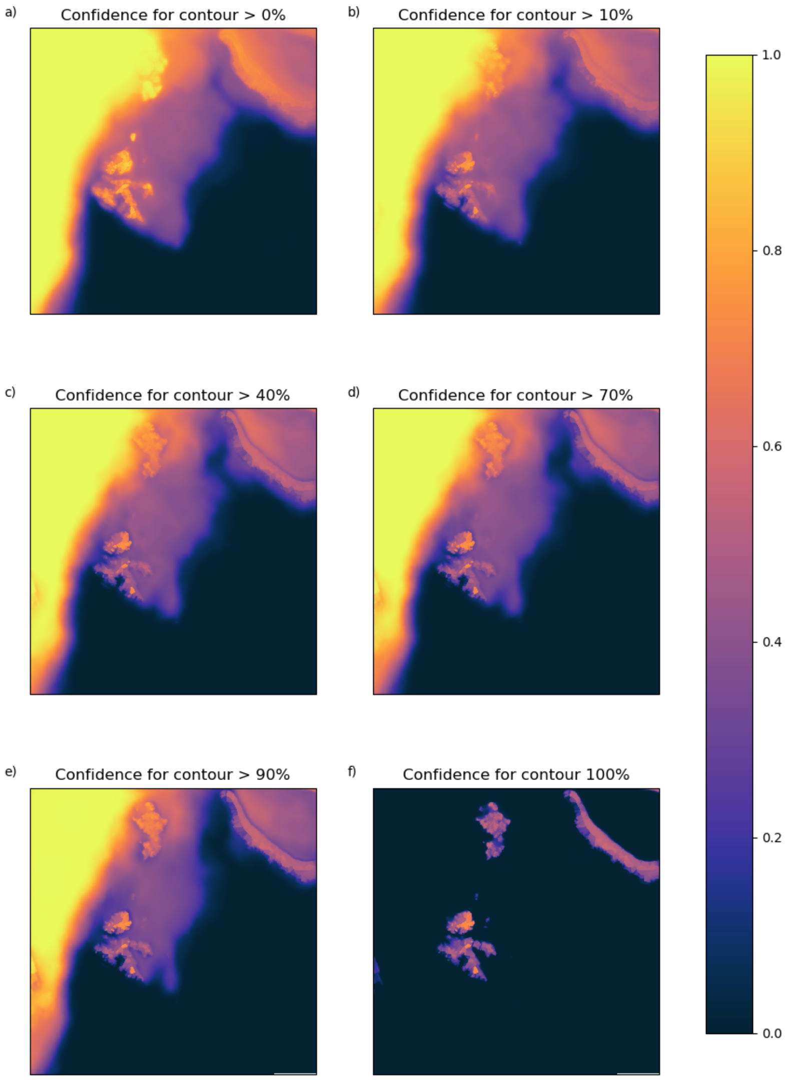
This section covers results regarding the multi-product comparison. First, the preparation of samples as well as setup of comparison environment is described. The physical models considered for this comparison are neXtSIM (Williams et al., 2021) presented in section 1.3.2 and Barents-2.5 (Röhrs et al., 2022) presented in section 1.3.3, whereas the considered baselines are persistence and the linear sea ice concentration trend described in section 1.1.

When comparing against multiple products, the coarsest resolution model is used as a common spatial resolution, although all products are interpolated onto the AROME Arctic projection. As both baselines have a daily forecast frequency, comparing either with a deep learning prediction involves identifying the forecast with similar bulletin date and valid date, i.e. initialized at the same day and targeting the same lead time.

Comparing against the two physical models are less intuitive, as both physical models have an hourly forecast frequency (Williams et al., 2021; Röhrs et al., 2022). 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 predictions made by the physical model when it covers the valid date of the deep learning forecast. 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 and the mean of the third predicted day is compared against a deep learning system with three day lead time. Figure 4 summarizes the process. Note that Barents-2.5 only have a 66 hour lead time (Röhrs et al., 2022), thus the mean between $t = 48$ and $t = 66$ is computed when comparing against a three day lead time prediction.

1.4 Preparing data

The logic behind sample creation is similar for both physical models. The idea is that the bulletin date of the physical forecasting system is $+1$ the bulletin date of the machine learning forecast. Furthermore, a daily mean is computed from the forecast based on the lead time of the forecast. I.e., a 1 day lead time for the machine learning forecast would



a)
Confidence for contour > 0%
b)
Confidence for contour > 10%
c)
Confidence for contour > 40%
d)
Confidence for contour > 70%
e)
Confidence for contour > 90%
f)
Confidence for contour 100%
1.0
0.8
0.6
0.4
0.2
0.0

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

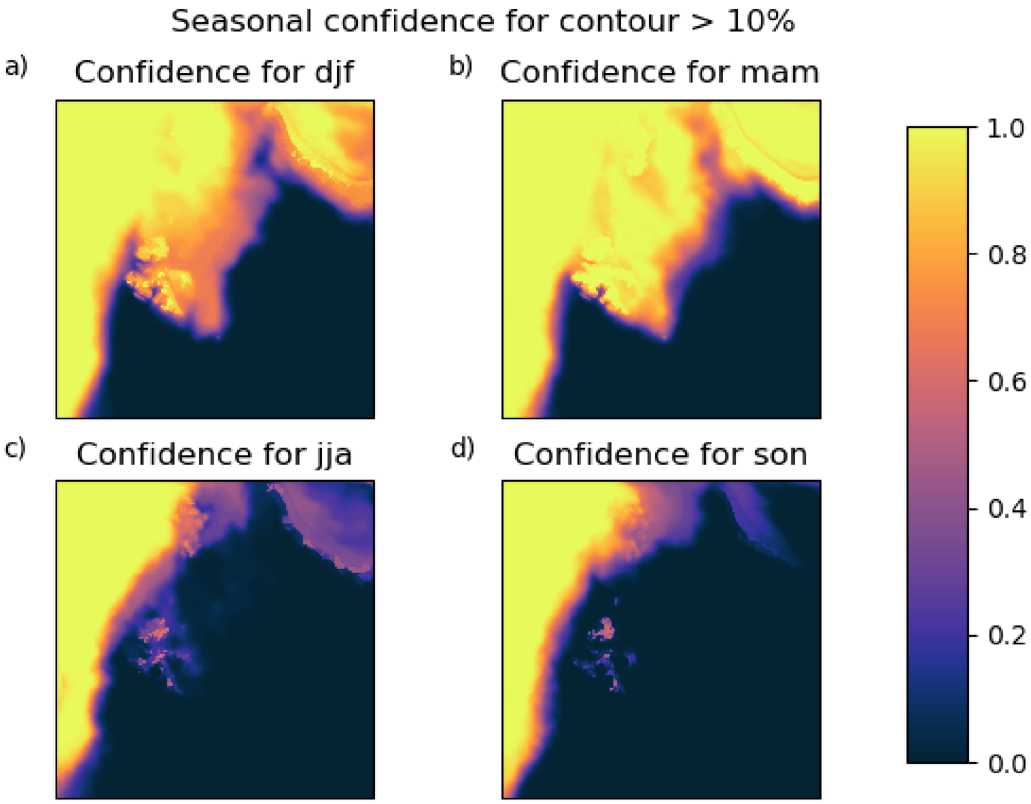


Figure 3: Mean seasonal confidence for the (> 10%) cumulative contour.

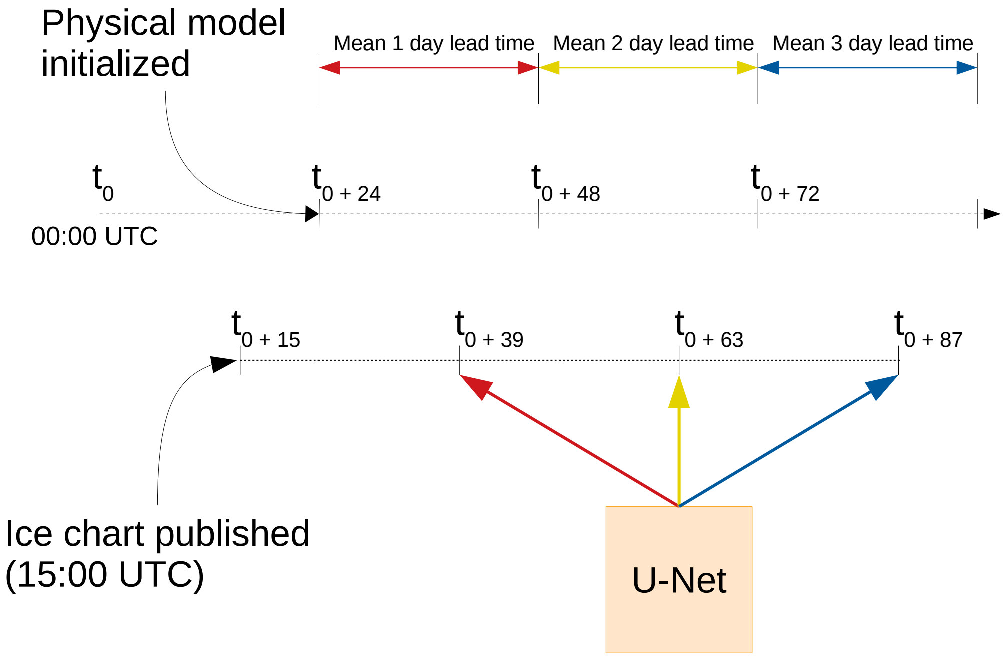


Figure 4: Overview describing how a physical model with an hourly frequency is compared against a deep learning forecast. Timestamps are hourly, and relative to 00:00 UTC the day a sea ice chart is published. The physical model is initialized the following day. Colors are used to denote comparability, with red = 1, yellow = 2 and green = 3 day lead time.

constitute a daily mean of the first 24 hours forecasted by a physical forecasting system starting at 00 the following day of the machine learning bulletin date.

References

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.: ”in prep for GMD” An operational data-assimilative coupled ocean and sea ice ensembleprediction model for the Barents Sea and Svalbard, p. 20, 2022.

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.