

# 1 Discussion

The following section will discuss the results presented in sections 1, 5 and 6. The results will be discussed separately and in conjunction where appropriate.

## 1.1 Initial attempt

Section 1.3.2 presents the initial attempt at designing a deep learning architecture for sea ice category classification. The initial architecture closely resembled the U-Net architecture as described by Ronneberger et al. (2015), and contains a single output layer with multiple channels such that multiple classes is predicted at the same time. Following the U-Net architecture and training procedure descriptions given in sections 1.3 and 1.4, where each channel in the output layer is representing a target class with the channels containing unactivated probabilities (logits) for each pixel to belong to the class. To determine which category each pixel is most likely to belong to, the softmax function (equation 6) converts the logits to probabilities where for each pixel is assigned the category belonging to the channel with highest probability. As seen in figure 5, neither the (40 – 60%) nor (70 – 80%) categories are resolved. This result (not shown) occurs for all samples in the 2022 test data. It appeared from the monthly distribution of sea ice categories presented in figure 3 that the intermediate sea ice categories (very open drift ice, open drift ice, close drift ice) constituted a significantly low fraction of the different sea ice categories. Due to the skewness in the sea ice category distribution towards ice free open water and very close drift seen in the sea ice charts, the current formulation of the segmentation task is highly imbalanced in favour of the aforementioned classes. Furthermore, due to the presence of class imbalance, it is expected that the computed loss (which is the unweighted variant of 8) is dominated by the more likely classes. Assuming that each contribution to the loss is treated equally due to the absence of a class- or pixel level weight term, the more likely classes supply a larger fraction of the learning signal which as shown in Lin et al. (2017) has the potential to dwarf or counteract the portion of the learning signal computed from the less likely classes.

With regards to the intended operationality and usefulness of the developed deep learning system, resolving the MIZ is a crucial aspect to achieve skillful forecasts that ensure maritime safety (Wagner et al., 2020). Considering the above discussion, it is likely that the model is not predicting the intermediate MIZ classes due to the model consistently predicting the larger area-fraction classes as more likely as a consequence of the class-imbalance (figure 3). One way to approach the problem of class imbalance would be to modify the weight parameter  $w$  in equation 8. In the work of Ronneberger et al. (2015),  $w$  was defined as a precomputed weight map which assigned some pixels more importance than others to compensate for the different frequency of pixels between the classes. However, it is noted

that the weight map computed by (Ronneberger et al., 2015) is specifically designed for biomedical image segmentation, and not directly transferable to other domains. Another approach to which indirectly defines the weight  $w$  is through rewriting the loss function, such as the focal loss proposed by Lin et al. (2017). The focal loss introduces a focusing parameter and a modulating factor which down-weights the loss contribution from predictions with a probability, i.e. easy to predict samples (Lin et al., 2017). Another way to approach the problem of class imbalance is to reformulate the prediction task to a more balanced problem as described in section 1.1.5 with the introduction of cumulative contours. Due to the immediate improvement in contour resolve seen when implementing the cumulative contours, the approach was further pursued in favour of the single output layer model.

## 1.2 Determining the depth of the model

Finding the optimal learning rate and depth of the desired deep learning architecture (multiple outputs, single label) is summarized in figures 6 and 7. Based on the conducted grid search in figure 6, we see that the validation loss tend to increase when the model becomes deeper, as well as when deviating from  $lr = 0.001$ . Although the lowest validation loss is achieved with  $lr = 0.001$  and a depth of 512 channels in the bottleneck, the difference is marginal compared to the 256 depth counterpart. However, the model with a depth of 256 channels has 4 times less parameters than the 512 depth model, which indicates that the 256 model is satisfactory fit to the data without needing the additional parameters found in the deeper network. This is further indicated by figure 7, which show that the 256-depth model is marginally improved beyond the 10th epoch.

When comparing a prediction from a 1024 depth U-Net with a 256 depth U-Net (figure 8), the predictions are generally visually similar. The similarity can be further seen when comparing the difference in NIIEE, which is only 2.5km between the two models. As such, the increased complexity gained from increasing the parameter-count of the model does not seem necessary to increase the predictive skill of the model. Viewed in conjunction with figure 7, a possible explanation may be that the 256 model already rapidly fits the training data. Hence, the limited number of training samples does not necessitate a more complex architecture in terms of encoder depth.

Using equation 4 from Araujo et al. (2019), it was calculated that the bottleneck of the 256 encoder have a theoretical receptive field of 145 pixels in each direction, whereas the 1024 encoder have a receptive field covering the entirety of the input fields. As such, each high level feature in the 256 bottleneck have only been influenced by lower level features in a 145km radius, whereas each high level feature in the bottleneck of the 1024 model have been influenced by features from the entirety of the domain. Hence, although the encoded signals in the 1024 bottleneck contain influences from the entire input field, figure

6 and 8 indicate that the performance is still reduced, as the model has a higher NIIEE compared to the 256 model with a limited influential range for the bottleneck.

It is shown in Luo et al. (2017) that the effective receptive field is limited compared to the theoretical receptive field, with the effective receptive field attaining an asymptotic gaussian shape. Luo et al. (2017) also show an example where although the theoretical receptive field is bigger than the input size, the effective receptive field is not able to fit the whole image, which is a consequence of the relative shrink of the effective receptive field shown to follow the relationship  $\propto \frac{1}{\sqrt{N}}$  where  $N$  is the number of layers (Luo et al., 2017). Thus, based on the results of Luo et al. (2017), all features of the bottleneck in the 1024 model are unlikely to be influenced by the entirety of the domain despite the coverage implied by the theoretical receptive field. Moreover, the full domain theoretical receptive field coverage was attained by increasing model complexity by a factor of  $\sim 16$ . Thus, the current results seem to discourage the need of fitting deep and complex deep neural networks to high resolution and spatially dependent data, as their theoretical ability to encode high level features influenced by all grid cells from the input data is opposed by a combination of the increased model complexity and reduced effective receptive field fraction.

### 1.3 Demonstrating seasonality

The deep learning system is able to preserve seasonality, as indicated by figure 9 which show that the predicted sea ice concentration for a two day lead time follows the observed sea ice edge for all months. Based on the demonstrated predictive capabilities of the model seen in figure 9, it can be seen that the deep learning system is able recreate the seasonal variability of the sea ice edge. Furthermore, the model is able to follow the seasonal cycle of sea ice concentration from the physical predictors only, which indicates that the model is able to infer the day of year based on the combined state of the sea ice concentration and atmosphere. This contrasts the work of Grigoryev et al. (2022), where the date was explicitly given to the model as a predictor. Thus, the model is able to capture the seasonal differences of sea ice dynamics from physical input fields only.

### 1.4 Using NIIEE as a metric

Throughout this thesis, model performance have been measured according to the normalized sea ice edge displacement error (Goessling et al., 2016; Melsom et al., 2019; Palerme et al., 2019; Zampieri et al., 2019) for different contours. The works of Goessling et al. (2016); Palerme et al. (2019); Zampieri et al. (2019) all applies the IIEE (or the Spatial Probability Score (Goessling and Jung, 2018), which is a probabilistic variation of the

IIEE) to seasonal prediction systems with coarse spatial resolution ( $\gg 1\text{km}$ ). Given that the IIEE is proportional to the number- and size of the grid cells, it is explored how the IIEE responds when used in a high spatial resolution sea ice concentration context. Moreover, the validity of deriving the NIIEE from a high resolution IIEE with a coarser resolution sea ice edge length is also inspected in section 1.3.1. Firstly, based on the computed correlations between the NIIEE computed from both sic and sea ice edge at 1km and NIIEE where both fields were at a 10km resolution, the NIIEE is conserved at increasing resolutions. Moreover, kilometer-scale resolution is not a threshold resolution which causes the IIEE to diverge.

Secondly, when dividing the IIEE from a 1km spatial resolution grid by a sea ice edge length computed from a 10km sea ice concentration field, the variability of the NIIEE is similar as when the 1km IIEE is divided by a 1km sea ice edge length (figure 4). The results indicate that the IIEE is a relevant metric also when applied to high spatial resolutions. Moreover, the variability of the NIIEE is preserved with increasing sea ice edge resolutions. Note that the sea ice edges used in section 1.3.1 were not independent of the sea ice concentration, as is the case with the sea ice edge length derived from OSI SAF CDR.

With regards to model selection, it was shown in section 1.3.5 and figure 14 that the correlation between the NIIEE and validation loss is strong with a value of 0.82. Moreover, figure 14 presents an example training where after each epoch, the model was measured in terms of the NIIEE on the validation set. As validating the deep learning system against the NIIEE was considerably slower (2 hours) compared to validating against the validation loss (8 minutes) for a single epoch, it was decided to select models in terms of the validation loss since the correlation between the metrics is strong.

## 1.5 Increasing the size of the training data

The size of the core training data (2019 and 2020) of 390, 289 or 286 samples for one, two and three day lead time respectively is considerably small when training an encoder-decoder deep learning model. Furthermore, the strong autocorrelation seen in the sea ice charts (figure 4) may act to further reduce the total training dataset, as the sea ice concentration for consecutive days have a low variability (Fritzner et al., 2020). However, by increasing the predictor pool through introducing atmospheric variables from AROME Arctic, the deep learning model is also learning correlations between the sea ice concentration development and atmospheric (temperature and wind fields) development. Hence, the strong autocorrelation seen in the sea ice concentration (figure 4) may not cause the unwanted effect of rendering certain samples redundant, as the atmospheric predictors allows the model to learn a more complex pattern from the input data.

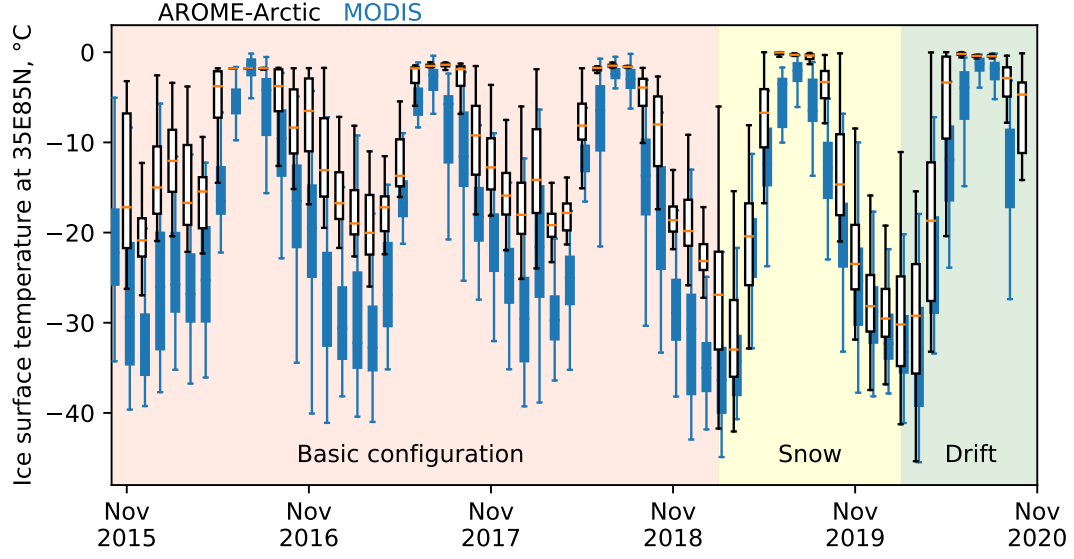


Figure 1: Ice surface temperature from AROME-Arctic compared with observations. Arome Arctic is distributed as white boxes, whereas observations are blue boxes. The snow on ice parameterization was added starting at the yellow background. Figure courtesy of Yurii Batrak.

It was mentioned in section 1.1.6 that the core training data only consists of the year 2019 and 2020 due to a major update regarding the representation of near surface temperature. It is common to assume that all data used to train, test and validate a machine learning algorithm is independently and identically distributed (IID), but this assumption does not hold when using predictors from AROME Arctic which has been continuously updated after it was released (Müller et al., 2017). The choice of limiting the training dataset to the aforementioned two year period was made based on an assessment of which updates would have a strong influence on the IID assumption for the chosen variables, where the snow on ice parameterization was deemed particularly significant (Batrak and Müller, 2019). An example of how the updated changed the near surface temperature distribution can be seen in figure 1, where a clear bias reduction compared to observations occurs following the update.

However, with regards to the problems occurring as a consequence of limited data, the effect of training on an increasing amount of years was shown in figure 11. From figure 11, it can be seen that both the NIIEE and validation loss is reduced when 2017 and 2018 is prepended to the training dataset, although 2016 causes the model to lose skill. Figure 12 show that the sea ice charts for the years covered by the training, validation and test dataset follow a similar distribution. Furthermore, there are no trends or clear outlier between the years. As such, it can be assumed that the sea ice charts are IID also when

the training data has additional years prepended. Thus, the loss of skill caused by 2016 is likely to be caused by predictors from AROME Arctic which, due to continuous model development, are not IID. That the loss of skill only happens with 2016 included could be a reflection of 2016 being the first full year of operation for AROME Arctic, and it is expected that significant updates occurred frequently during AROME Arctic's first year of operation.

It should also be noted that the inclusion of 2017 and 2018 seems to improve model performance, despite the years not including the snow on ice parameterization deemed significant with regards to the representation of t2m. This could be explained by the model being more impacted by the winds rather than the temperature from AROME Arctic, thus causing the shift in temperature variance to be negligible. A further discussion of predictor importance and model explainability will be done in section **NOT YET**.

Figure 11 highlights the difficulties of fitting a deep learning model to data from an operational product, as the data is updated frequently in order to improve the product performance through intentional shifts in model bias or variance (Batrak and Müller, 2019) (see also figure 1). This highlights that there are fundamental conflicting interests in terms of dataset consistency for operational numerical weather prediction output data and data used to train a deep learning network.

It was mentioned in section 1.3.1 that re-forecasts or reanalysis products such as CARRA (Køltzow et al., 2022) or ERA5 (Hersbach et al., 2020) would alleviate the problem of temporal inconsistencies in the data, as the model used to create a reanalysis is based on a single model cycle. However, reanalysis products generally have a publishing delay from days to months, which is not compatible with the operational timeliness of providing forecasts on the same day as the predictors are valid envisioned for the developed deep learning system. Furthermore, a reanalysis differs from a forecast in that data is complete and consistent, which in terms of the temporal mean approach for the atmospheric predictors would limit the data diversity as the reanalysis provide the same values for a given date regardless of the initialization date of the deep learning forecast. This is contrasted by a forecast, which output different values at shared timestamps depending on the forecast initialization date. Hence, despite the expected increase in skill of a reanalysis product, the above may serve to limit the overall usefulness for reanalysis products as deep learning predictors in the case where there is a temporal overlap between predictors. On the other hand, re-forecasts as a static model cycle providing full range forecasts would not have the same restraints since predictors with overlapping timestamps would be computed from different values.

Although this project have attempted to increase the dataset size directly by adding more data, similar studies utilized other techniques such as Grigoryev et al. (2022) which trained their deep learning model on multiple regions and performed geometric transformations on

the data during training. Their results show that geometric augmentations have a small positive influence for some areas, and negligible for one area. Though adding multiple regions to increase dataset size is outside the boundaries of this work, applying geometric augmentations to the data is a known generalization technique for image based deep learning which Grigoryev et al. (2022) could alleviate the limited dataset if implemented correctly (Shorten and Khoshgoftaar, 2019). However, given that atmospheric variables have a strong spatial dependency inferred by their lat/lon coordinates, it is questioned to what extent geometric transformations applied to the entirety of a scene can teach the network new relations between the variables.

## 1.6 Model behavior from tuning hyperparameters

The effect of interpolating sea ice concentration over land covered pixels following the approach of Wang et al. (2017) was recorded in section 1.3.4. The approach was compared with assigning all land-covered pixels as the category ice free open water (open-ocean mask), where the latter approach increased the mean annual NIIEE with 1.5km. This shows that how the sea ice is represented in areas which are ignored during validation due to the land sea mask impacts how surrounding grid cells are classified. Moreover, the result show that supplying the land-sea mask as a predictor is not by itself sufficient to teach the network to discern between land pixels and water pixels, which may be explained by how the convolutional kernel computes features from a local neighborhood separately for each channel (Fukushima, 1980).

Another interpretation of the open-ocean mask is that it relates the affected predictor to be treated as an incomplete partial convolution (Liu et al., 2018). Partial convolution is a variation of the convolutional layer which stem from image inpainting, which is the task of filling holes in an image (Liu et al., 2018). The technique involves multiplying the convolutional filter with a masked variation of the local neighborhood determined by the holes in the image, then the computed value is scaled to adjust for the number of unmasked pixels in the neighborhood (Liu et al., 2018). Since the integration of the land-sea mask into a predictor through the open-ocean mask was without a consideration of scaling the output with regards to the varying number of masked pixels in the local neighborhood, the sharp gradient computed between land-fast ice and ice-free open water is left unscaled thus detected as a notable feature. Thus, the work of Liu et al. (2018) may provide a possible explanation as to why the open ocean mask degrades the performance of the model.

However, the result also highlights a different approach to integrate the land-sea mask into the predictors, as partial convolutions would make it possible to mask a subset or all predictors with the land sea mask thus removing it from the pool of predictors. This may aid to increase the relevance of the land-sea mask, as although the intended behavior

of the predictor is help the model treat land pixels differently than ocean pixels it was previously discussed that this behavior was not directly inferred from the land-sea mask. By utilizing partial convolutions, and using the land-sea mask as a mask, the convolutional result would only depend on the non-hole regions (Liu et al., 2018).

Reducing the number of classes to output does not seem to improve the model beyond the baseline. The model trained in section 1.3.4 where the  $> 0\%$  and  $= 100\%$  contours was removed performs similarly to the baseline model which outputs all target classes, with an increase of the NIIEE of 0.4km. This result may suggest that although removing contours have a theoretical impact on training, the effect is not reflected by the predictions. From the description of how the individual losses computed with equation 4 are treated as a sum starting at the decoder during backpropagation (section 1.2.7), it can be seen that removing contours decreases the magnitude of the loss function. Moreover, due to the cumulative contours formulation, removing outputs also has the effect of lowering the amount of redundant information as the number of overlapping pixels is reduced. Hence, the result of removing contours, especially the 100% contour, is that the computed loss is less weighted by the land-fast ice pixels

The mean seasonal confidence fields from figure 31 can be used in conjunction with the loss function defined in equation 4 to determine the impact of the  $= 100\%$  contour have on the overall loss starting from the decoder. Equation 4 is defined such that a greater deviation from the true value (1) results in a greater loss. Moreover, the  $= 100\%$  contour is usually present beneath the land areas as the fast-ice contour drawn close to land by the sea ice specialist is interpolated onto the land-pixels by the nearest neighbor interpolation (Wang et al., 2017) performed on the sea ice charts during data preprocessing (section 1.1.2). Figure 31 reveal that the  $= 100\%$  contour has a seasonal contribution to the loss, where the winter and spring contribution is lower than the summer and autumn contribution. Since the confidence levels are high during winter and spring, the impact from the  $= 100\%$  contour expected to be small. However, the overall low confidence levels during summer and autumn will increase the loss hence samples from the summer and autumn seasons trained on networks where the  $= 100\%$  contour is used as an output may decrease model performance if the loss from the other contours are dominated by the loss of the  $= 100\%$  contour. As it is ideal for the model to perform well at resolving the MIZ related contours, reducing the number of contours such that unrelated contours with a high loss (such as just described for the  $= 100\%$  contour) does not negatively impact model performance through high influence on the loss is advisable.

Another possible way to mitigate the loss from one cumulative contour to dominate the loss contributions from the other cumulative contours is to reformulate the network architecture as a multitask learning problem. Multitask learning is a branch of machine learning where a single network is used to perform multiple tasks simultaneously, with the goal of sharing as much information as possible while at the same time reducing neg-



ative interactions between the tasks (Crawshaw, 2020). For the current task, a negative interaction could be the aforementioned high loss from the ( $= 100\%$ ) contour which decrease the ability for the other contours to influence the loss when training for certain seasons. Multitask networks for computer vision tasks generally employ an approach originally proposed by Zhang et al. (2014) where a common feature extractor is shared by individual output branches for each task (Crawshaw, 2020). This general approach can be translated to the U-Net architecture, where the encoder is used as a common feature extractor and each task is assigned a individual decoder (Jha et al., 2020). Given the cumulative contour formulation, employing a multitask learning U-Net architecture similar to Jha et al. (2020) where each contour is predicted individually with its own decoder could help differentiating the predicted contours as they do not share the same pathway, as well as reduce negative interactions as described above. However, this approach could cause the network to disassociate the cumulative contours from one another, causing unwanted effects on the output forecast such as a higher amount of sporadic category change which is currently not the case for the shared decoder models. E.g. both figure 8 and 9 show the contours tend to gradually change starting from the lowermost contour, with few occurrences of sharp category changes.

The model response to replacing all ReLU non-linear activation functions (Nair and Hinton, 2010) with a linear mapping was visually exemplified in figure 13. The linear model increased the mean annual NIIEE from the baseline by 13.15km, which is 6km higher than persistence mean annual NIIEE for the test data as shown in figure 41. The purpose of assessing the forecast skill with a linear deep learning system is to understand the effect of the non-linear activation function. We see that predictions made with a linear model perform significantly worse than persistence (for a two day lead time), which renders the forecasts skillless. The example prediction seen in figure 13 may provide insight as to why the linear model performs worse than the non-linear counterpart. Firstly, there are several instances of checkerboard artefacts commonly caused by deconvolutional layers. The U-Net architecture contains skip connections from the encoder to the corresponding layer of the decoder (see figure 2, gray arrow) where spatial information in the encoder is concatenated with the upsampled signal from the deconvolutional layer and may help to suppress the checkerboard artefact (Ronneberger et al., 2015). However, the two concatenated feature maps are not merged until after the first convolutional layer and subsequent activation function. Hence, when the non-linear activation is replaced by a linear mapping, the convolutional layers are not able to suppress the checkerboard artefact by themselves, which shows that the non-linear connections introduced with the ReLU activation function are crucial for model performance.

Secondly, the linear model seem to underestimate the  $\geq 10\%$  contour in favour of the  $\geq 40\%$  contour when compared to the same prediction with the baseline model in figure 8 (a). The linear model is also unable to resolve the isolated patch  $\geq 70\%$  sea ice

concentration North of Novaya Semlya seen in figure 8. Moreover, detailed structures such as the what is seen for the  $\geq 90\%$  contour towards the lower left of figure 13 are likely remnants of the sea ice chart used as predictor. Thus, the model is repurposing parts of the input when making a prediction, which is probably done due to the high autocorrelation seen between the sea ice charts (figure 4). However, this means that the linear model have learned limited connections between the predictors, indicating that the non-linear activation function is necessary for the model to correlate sea ice dynamics with the state of the atmosphere as discussed in section 1.3.

## 1.7 Performance

As described in section 5.1, a forecast is skillful if it achieves a lower NIIEE than persistence. The performance of the developed deep learning system was measured in two ways. Firstly, the performance of the deep learning system was measured against persistence, utilizing the sea ice charts as the ground truth target. Secondly, the performance of the deep learning system was compared against four other sea ice concentration products, with two different ground truth datasets utilized where one of them (AMSR2) is completely independent of the deep learning system to measure the generalizability of the product.

### 1.7.1 Model performance with a two day lead time

The 256 architecture forecasting with a two day lead time was compared against persistence, computing the NIIEE for all contours. The result was presented in figure 29, where it was shown that the deep learning forecast achieves a lower median and interquartile range boundaries for the  $\geq (10, 40, 70, 90)\%$ . The model tend to perform significantly worse than persistence for the fast-ice contour (figure 29 (f)), which can be explained by the limited fraction of the scene which contains fast-ice (figure 3) causing the scene to be skewed in disfavour of the category even with the cumulative contour definition. This is further elaborated through figures 30 (f) and 31, which show that the model predicts a small fraction of the domain as the contour, and that the confidence of the prediction is tied to seasonality. Furthermore, inspecting the other confidence fields in figure 30 reveal that the model predict the other categories with higher probabilities and with a strong resemblance to each other. Thus, when each contribution to the total loss is summed at the end of the decoder during backpropagation, the similar signals from the intermediate contours increases the magnitude of the computed loss in the shared decoder.

For the  $> 0\%$  contour, the model is performing worse than persistence in the summer, with the summer distribution for the deep learning system closely resembling the persistence

distribution although it appears somewhat denser around the 25th percentile. Moreover, it is seen that the deep learning forecasts have a smaller shift in interquartile range during the summer and autumn compared to winter and spring for all contours (except  $= 100\%$ ). This may in part be explained through the NIEE, which is proportional to the sea ice edge length Goessling et al. (2016) as the potential area of error is reduced. However, it was discussed in section 1.6 that as a result of not utilizing a multitask approach to the U-Net architecture, the  $= 100\%$  contour could negatively impact training due to potential high errors over the land-fast ice contour dominating the computed loss. The apparent seasonality in model performance seen in figure 29 could be explained by the negative interactions between the  $= 100\%$  as it has low confidence which causes high errors during summer and autumn. Following the discussion in section 1.6, that the  $= 100\%$  has a different confidence spatial extent compared to the other cumulative contours (figure 31) and especially during winter and autumn where the fast-ice contour is generally disconnected from the other contours. Thus, the negative impact caused when the different contours interact in the shared encoder decreases model performance during summer and autumn as seen in figure 29. This may not be a problem during winter and spring, as the  $= 100\%$  contour is not disconnected from the other cumulative contours, hence not contributing with loss values irrelevant to the extent of the sea ice edge.

Figure 29 demonstrates the potential for machine learning forecasts of sea ice concentration, as the deep learning system generally outperforms persistence. The relevance towards maritime end users is encouraged by the performance at the lower concentration contours (such as the  $\geq 10\%$ ) (Wagner et al., 2020; Veland et al., 2021). Moreover, the results demonstrate the effectiveness of formulating the targets as cumulative contours rather than multiclass classification which was initially discussed in section 1.1. The cumulative contours tend to sufficiently resolve the intermediate sea ice categories (MIZ). It appears that the intermediate classes are well resolved at the expense of the fast-ice contour with the cumulative contour formulation, as seen in figure 33. With respect to end user relevance, this is an apt tradeoff.

### 1.7.2 Model performance for varying lead times

It is shown in figure 10 that model performance decreases in terms of absolute NIEE as a function of lead time. However, the relative improvement over persistence increases with lead time, as also shown in figure 10. Based on the results of Zampieri et al. (2019), it is expected that the NIEE for persistence increase fast with lead time. However, the results from Zampieri et al. (2019) also show that no seasonal sea ice prediction system has predictive skill for the first few ( $\sim 2$ ) lead times, indicating that persistence forecasts are relatively skillful for short lead times. Hence, achieving lower NIEE than persistence for all considered lead times further demonstrates the predictive capabilities of the deep

learning system. The relative improvement over persistence seen in figure 10 increases with lead time, and can be attributed to the rapid loss of skill for persistence forecasts as seen in (Zampieri et al., 2019). It is noted that the error in figure 10 are computed from three individually fitted models, where the atmospheric predictors are covering a larger temporal range as function of lead time (section 1.1.4). Thus, the error at different lead times in figure 10 are independent of each other. This shows that the predicting sea ice concentration at increasing lead times also becomes increasingly difficult, despite modifying certain predictors to conform to the increased lead time.

Figure 32 provide some insight into how the different deep learning systems resolve sea ice concentration contours by showing the mean monthly sea ice edge length. From figure 32, it can be seen that the sea ice edge length decreases as a function of lead time. Moreover, the figure already shows that there is a negative bias in terms of sea ice edge length when comparing the forecasts with the target sea ice chart. The negative bias in sea ice edge length can be explained when comparing figures 8 or 9 with an example sea ice chart as presented in figure 2, where it can be seen that the predicted sea ice charts have a smoother appearance rather than the sea ice chart which resolve individual structures with a high level of detail. As the sea ice edge length is computed as a sum of grid cells (see equation 12) (Melsom et al., 2019), it follows that a less detailed sea ice edge results in a shorter sea ice edge length. Since the ice edge length becomes shorter for increasing lead times, it is expected that the deep learning forecasts produce higher variance forecasts which are less precise, i.e. smoother contours.

### 1.7.3 Comparing against multiple products

A difference between the two comparisons in sections 5.2 and 5.3 is the choice of grid, where the AROME Arctic grid native to the sea ice charts and deep learning system used in section 5.2 is exchanged with the grid of the product with largest spatial resolution in section 5.3. Hence, when utilizing the sea ice charts as ground truth, the 3km neXtSIM grid is utilized (Williams et al., 2021). When AMSR2 is used as ground truth, the 6.25km AMSR2 grid is utilized (Spreen et al., 2008). Since nearest neighbor interpolation is used to downsample the products onto the target grid, it is expected that the higher resolution products such as the deep learning system or persistence receive a larger amount of interpolation artefacts compared to the products with a grid which is closer to the target. Although the interpolation artefacts remain unmeasured, it is assumed that they have a negative effect proportional to the grid size difference on the products.

The inter-product comparison made between NeXtSIM, persistence, the baseline deep learning model, the linear OSI SAF trend and Barents-2.5 for a two day lead time was shown in figure 35. It is initially noted in section 5.3 that only the sea ice chart based products are able to resolve the  $> 0\%$  contour. When comparing data against the  $> 0\%$

contour, it is important to note that the contour itself is not resolving the sea ice edge, nor is it directly tied to any observation of sea ice. However, the contour is drawn by the sea ice analyst to denote a buffer tracing the actually observed sea ice edge denoted by the  $\geq 10\%$  contour. Thus, the  $> 0\%$  contour is a result of the operational nature of the sea ice charts, rather than a actual occurrence. This result is expected, since the other sea ice concentration products used in the inter-comparison are independent of the sea ice charts. With the OSI SAF linear trend being computed directly from satellite observations (Tonboe et al., 2017) and both physical models being forced by the ocean and sea ice model TOPAZ (Sakov et al., 2012).

Similarly to when the deep learning system was compared only against persistence in section 5.2 figure 29, the deep learning system exceeds all other products in terms of distribution mean and median for the  $\geq (10, 40, 70, 90)\%$  contours, further demonstrating the skill of the deep learning forecasts. Figure 35 also provide insight into the performance of the different physical sea ice prediction system, where Barents-2.5 tend to provide lower median and mean values compared to NeXtSIM. Moreover, neXtSIM tend to increase the distribution spread for increasing contours, especially during the summer season. The OSI SAF linear trend is performing worse than persistence as a baseline product, and the OSI SAF linear trend NIEEE distribution tend to increase with increased target contour. The lack of performance for the OSI SAF linear trend can be connected to the subpar performance of the linear deep learning model discussed in section 1.6, and indicates that the sea ice concentration dynamics are of a non-linear nature (Grigoryev et al., 2022).

The spatial anomalies for all products used in the inter-product comparison with respect to the sea ice chart targets was shown in figure 37. The overall good performance for both the sea ice charts as well as persistence is reflected in figure 37 since the biases are overall low. Compared to all the other products, the sea ice chart based forecasts have no bias in the upper left corner for all seasons. This corresponds with where the deep learning forecast always showed high confidence for predictions, further indicating the aforementioned feedback in that the model is apt at predicting where the cumulative targets overlap due to the shared decoder. Furthermore, figure 37 provide some insight into the performance of neXtSIM, as there is a consistent negative bias along the sea ice edge which is especially prominent during winter and spring. Figure 37 show that neXtSIM underestimates the sea ice edge during winter and spring, and the entire scene for summer and autumn. Which may explain why neXtSIM performs the way it does.

The anomalies for Barents-2.5 are generally low, but covers large parts of the available scenes. The consistent underestimation of sea ice concentration seen in Barents, neXtSIM and to some degree the OSI SAF linear trend in figure 37 during summer and autumn may occur due to the physical models assigning the grid-cells concentration values less than 90%, whereas the  $\geq 90\%$  sea ice category is frequently used by the sea ice specialist to denote all sea ice that is not part of an ice edge. This highlights potential weakness

in comparing the physical models against sea ice charts, and may also explain why the physical models are less comparable than the sea ice chart products in figure 35 as it seems the non-sea ice chart products forecast lower concentration values, especially during summer and autumn.

With respect to operational concerns of the developed deep learning system, as well as with regards to providing user-relevant validation metrics which are easily interpretable (Veland et al., 2021), the fraction of days where the deep learning forecast offers an improvement compared to the other products was shown in figure 36. Firstly, it should be noted that the high uncertainty for months where the fraction of days is less than 100% arise due to limited monthly sample sizes (table 1). However, figure 36 is consistent with the performance for varying lead times shown in figure 10 when comparing the deep learning system against persistence. It can be seen in the top row of figures in figure 36 that the improvement against neXtSIM is decreasing with increasing lead time, which may indicate that neXtSIM gains performance with lead time.

A similar trend occurs when comparing against Barents-2.5, however the Barents forecast is able to achieve more days with lower NIIEE for November and December with a two day lead time. The consistent high performance of Barents seen in figure 35 and 37 motivate for an inclusion of the model into the pool of predictors. Barents-2.5 could provide forecasted sea ice concentration as a predictor similar to the atmospheric fields from AROME Arctic. Although Barents-2.5 is still in development, providing re-forecasts of the model as described in section 1.5 would allow for coverage of the training dataset with the current model performance as of after the spin up time of the data assimilation system (Röhrs et al., 2022). Furthermore, the problem regarding predictor multicollinearity as described in section 1.1.4 is avoided as Barents-2.5 provide sea ice concentration information independent of the sea ice charts, as well as providing data with a future perspective which could greatly enhance the skill of the deep learning system.

To assess the generalized performance of the deep learning system, the inter-product comparison was extended to also compare against independent AMSR-2 sea ice concentration observations (Spreen et al., 2008). Firstly, the sea ice chart based products are not skillful for the  $\geq 0\%$  contour, confirming the unphysical nature of the contour for the sea ice charts. However, for increasing contours starting from the  $\geq 40\%$  contour, the deep learning system consistently achieves the lowest median and mean, which means that the deep learning system is not biased towards only having skillful predictions for sea ice chart like sea ice concentration conditions. Since figure 38 show that the deep learning system outperforms the other products when targeted with data which has not been seen during training. Thus, the generalizability of the deep learning system is satisfactory. That the deep learning system is outperformed for the  $\geq 10\%$  may be a consequence of the uncertainty related to the AMSR-2 observations. Where it was described in section 1.2.5 that the ASI sea ice algorithm exerts higher uncertainties for lower concentration values,

mainly due to errors introduced through atmospheric interactions (Spreen et al., 2008). It is further noted that the AMSR-2 observations are most certain at 65%, thus the deep learning system achieves the lowest NIIIE distribution for the cumulative contours which correspond to where the target sea ice concentration values are most certain.

Finally, the model trained with reduced classes was not shown in figure 39 to improve the contours which it targeted in common with the baseline model used in figure 38. This is consistent with the results discussed in section 1.6 with the mean annual NIIIE of the reduced class model being similar to the baseline model. However, when not targeting the unphysical  $\geq 10\%$  contour, the model is able to achieve comparable performance akin to the other products independent from the sea ice charts, as seen in the top left of figure 39. Thus, to achieve model generalizability in terms of other sea ice concentration products than the sea ice charts, the  $\geq 10\%$  contour is advisable to remove as it hinders model performance at the contour.

## 1.8 Explainability

Two general approaches were used to determine the overall effect each predictor had on the deep learning system. The first approach was to leave one predictor out, and train the model with the remaining predictors. Although deep learning models fitted towards different pools of predictors are not directly comparable since a model is uniquely fitted to the available predictors, the overall skill of the differently fitted models can still be compared. The second approach involved modifying predictors in the test data set, such that model response to differently constructed out-of-distribution predictors could be measured. Out-of-distribution predictors signify that a input-channel or input combination is different than what was observed during training (DeVries and Taylor, 2018), and it is noted that all data processed by the model is considered out-of-distribution following the first convolutional layer where the input channels are merged into 64 feature maps.

For the current work, out-of-bounds distributions are reflected through replacing a predictor with uniform noise 42 or shuffling the sequence for only a predictor-variable while the other remain untouched 43. It is common for deep neural networks to provide incorrect predictions with a high level of confidence when fed meaningless predictors (DeVries and Taylor, 2018). However, given that the predictors for the current work are physically bound, although deemed nonsensical by the deep learning system the modification performed which causes the predictor to become out-of-sample can still be explained and in some cases related to physical processes. Moreover, given the limited size of the training data, it can be assumed that there exists a number of valid atmospheric configurations not seen by the deep learning system which to some extent can be replicated by the following experiments. Thus, a systematic model response to a consistent and explainable

predictor modification may provide insight into model response which is not just artificial hallucination.

### 1.8.1 Model response to predictors

It was shown in figure 41 that removing either the recent sea ice chart or all AROME Arctic variables caused the deep learning system to achieve worse performance than persistence. In terms of removing the recent sea ice chart, this result indicates that the recent sea ice chart is the most important predictor for the deep learning system. This is followed up by figures 42 and 43, where the model response to both modifications to the recent sea ice chart caused the model performance to significantly worsen. The high values when replacing the recent sea ice chart with uniform noise in figure 42 indicates that when the model is strongly fitted to the recent sea ice chart since the response is very much higher than for any other predictor. Comparably, the seasonal NIIEE is lower when the recent sea ice chart is swapped within the distribution, indicating that the model is able to recreate a sea ice edge when fed a actual sea ice chart. Hence, given that the model is strongly fitted to recent sea ice chart, it can be assumed that the predictions made by the deep learning system are in general modifications of the recent ice chart. This may be an effect of the strong autocorrelation inherent to the sea ice charts (figure 4). Such that when given a sea ice chart from a potential other season in figure 43, the forecast retains the seasonality inherent of predictor sea ice chart, which is not possible when the recent sea ice chart is replaced by uniform noise in figure 43. Thus explaining the large difference in seasonal NIIEE values for sea ice concentration in figures 42 and 43.

Figure 41 also shows that the deep learning forecast is not able to outperform persistence without atmospheric predictors from AROME Arctic. Thus, it seems to be the case that the deep learning system is not able to fit the presumed non-linear nature of sea ice dynamics (Grigoryev et al., 2022) without correlations from additional predictors such as the atmospheric conditions. Moreover, given the model responds well to relevant physical predictors, it can be assumed that further increasing the predictor pool with physical processes related to the MIZ may increase the skill of the forecast. Currently, the pool of predictors cover sea ice and atmospheric interactions, however the ice-ocean interactions present in the MIZ are missing. A driver of sea ice breakage in the MIZ is the interaction between the sea ice edge and surface waves, causing ice breakage which resonates with large scale dynamics and thermodynamical properties of the sea ice (Williams et al., 2013). Given that the model responds positively to physical predictors from the atmosphere, including relevant fields from a high resolution wave forecasting system such as Carrasco et al. (2022) where significant wave height and wave direction to be included as model predictors.

Figures 41, 42 and 43 also show that the model gains no skill from the linear sea ice



trend computed from the OSI SAF observations. In figure 41, this is shown as model performance is similar with and without the predictor. In figure 43 it is seen as the trend follows the baseline and exerts no variability when swapped with trends from other seasons. It can, however, be seen in figure 42 that replacing the trend with uniform noise improves some seasons, although stochastically. This may show that the predictor have a very slight degrading effect on the deep learning forecast, however it is more likely to be stochastic hallucination as the other figures 41 or 43 does not recreate a similar trend.

### 1.8.2 Model inferred physics

The same precautions regarding out-of-distribution samples given in section 1.8 also apply to the synthetic AROME Arctic samples explored in section 6.2, since the synthetic fields are likely to contain combinations of valid values which have not been covered by the training dataset. With respect to the forecast errors obtained when comparing a non-synthetic forecast with a forecast where one or several AROME Arctic fields have been replaced with a synthetic variation in figure 44, no clear seasonal cycle to the NIIEE can be seen. Such a random change in NIIEE for the same synthetic field at different seasons may show that the model is able to infer some relationship between the state of a physical field and the other predictors. This is further demonstrated by the two synthetic fields in the lower leftmost plot in figure 44 where the model improved the NIIEE when being fed no-winds or only positive winds in the x direction. Although there is possibility of these results occurring stochastically, it is noteworthy that there may be a possibility to engineer more ideal conditions if the state of other predictors are known.

These results in section 6.2 also indicate that the model does interpret each atmospheric field separately at the beginning of the encoder before the first convolutional layer merges the different input channels into a single feature map, since the model is able to evolve the sea ice concentration field differently given different states of the atmosphere. Additionally, when figure 44 is seen in relation to swapping the order of a single AROME Arctic field in figure 43, the general model response to a modified state of the atmosphere is to degrade the forecast skill. Which is expected given that the 2-meter temperature field and the wind fields are assumed to have a strong correlation, such that when a uncorrelated atmospheric field is introduced to the predictor pool the encoded signal is weaker resulting in less skillful forecasts.

The response to some of the synthetic AROME Arctic predictions in figure 44 was shown in figure 45 as spatial anomalies. The top row of figure 45 show model response to synthetic wind fields, whereas the bottom row show model response to synthetic 2-meter temperature fields. A major difference between the top and bottom row of figure 45 is where sea ice growth and decline occurs. For the top row, where only the winds are

synthetic, it appears that the spatial distribution of sea ice is located where sea ice already exists. However, for the 2-meter temperature synthetic fields in the bottom row, it is seen that the sea ice growth and decline occurs throughout the entirety of the scene, without relation to where the sea ice concentration is located in the recent sea ice chart predictor. This response is compliant with the expected physical interactions between the sea ice and atmosphere.

Firstly, the x and y-component of the wind is only able to affect sea ice dynamics for already existing sea ice (Spreen et al., 2011; Yu et al., 2020). From the upper leftmost plot in figure 45, it can be seen that then both wind components are directed in a negative direction, the extent of the sea ice edge is lowered. This is compliant with how the sea ice edge tend to be sharply defined when experiencing incoming winds, since the negative direction of the wind restricts the outward transportation of low concentration sea ice (Yu et al., 2020). The contrary seem to occur in the top right plot in figure 45, since the sea ice edge is relatively normal to the direction of the winds (only x wind in positive direction) causing the winds to transport sea ice away from the sea ice edge. This is reflected by some areas experiencing a wider MIZ, although with some inconsistencies which may be due to the positive x-wind not affecting the scene in isolation. Secondly, with synthetic 2-meter temperature fields the deep learning system is able to infer growth in areas where there are no sea ice in the recent sea ice chart similar to the occurrence of sea ice formation as a response to freezing temperatures (Hibler, 1979). Furthermore, the lower rightmost plot in figure 45 show a 2-meter temperature field which linearly increases from the lowest possible to the highest possible values in the test data from bottom to top of the scene. The sea ice concentration response to the synthetic forcing is similar, with growth in the bottom half of the domain and growth in the top half, as expected with regards to the physical response where cold temperatures facilitate sea ice growth and vice versa (Hibler, 1979). These results indicate that the deep learning system is able to infer physical relationships and responses from the predictors without having learnt them explicitly. Furthermore, the deep learning system is able to recreate physical responses between predictors to some degree, without having the framework to simulate or resolve the underlying physics.

## 1.9 Explainable predictions

This work opted to implemented the seg-GradCAM technique (Vinogradova et al., 2020) to increase the transparency and explainability of the developed deep learning system. An initial attempt implementing the technique for the baseline model with a two day lead time at different target contours was shown in figure 46. The visually highlighted region seem to follow the marked sea ice edge, although the regions in figure 46 (b, c and d) also seem to include parts of Svalbard to a greater extent than (a) showing that the model look

at different regions to determine the different contours. The only class each output layer is attempting to predict is whether each pixel belongs to the cumulative contour. Figure 46 show that the regions of the predictors which were important for predicting pixels as part of the cumulative contour were the pixels which constituted the predictor sea ice chart. Although for some contours the model is utilizing pixels outside the sea ice edge for the given contour, the region of the predictors which overlap with the non-highlighted portion of figure 46 were not important for predicting cumulative contours.

In section 1.2, it was shown that the shallow model without an scene-encompassing theoretical receptive field at the bottleneck outperformed the deeper model where each feature in the bottleneck had a receptive field which covered the bottleneck. Although it was discussed that the effective receptive field covered a fraction of the theoretical receptive field (Luo et al., 2017), based on figure 46 it can also be seen that by increasing the theoretical receptive field each encoded feature in the bottleneck are computed from a lot of features from unimportant spatial locations are included. On the contrary, a model such as the 256 U-Net architecture which only had a theoretical receptive field of 145 in the bottleneck would have a some features in the bottleneck only computed from important pixels, and some encoded features only computed from unimportant pixels. Thus, it may be the case that due to the clear dichotomy of low level features, the signal is easier to decode compared to when all features are a mix of important and unimportant pixels as for the 1028 U-Net architecture. Thus, figure 46 also indicate that it is not necessarily a positive for all features in the bottleneck to cover the entire scene, as not all pixels are equally important for predicting the cumulative contours hence diluting the features in the bottleneck with unimportant features.

A comparison is made between the activation maps from the baseline deep learning model in figure 46 and the activation maps from the model with reduced classes in figure 47. The immediate difference is that the highlighted regions in figure 47 are of much higher importance, compared to figure 46. However, it is noted that the baseline model have been used to create activation maps for different dates in figure 48 which show higher importance akin to the values in figure 47. It is not known if figure 46 contains unusually low values. Despite the general difference in strength between the activation maps in figure 46 and 47, it is noted that they differ in how they rank land pixels such as Frans Josef Land and Novaya Semlya in terms of importance. In figure 46 pixels related to Frans Josef Land is deemed very important for all contours, and this is also seen in figure 48 (a). It is also seen that Novaya Semlya and Svalbard achieve importance simultaneously starting at figure 46 (b). Although only one date is inspected, this may be related to the shared decoder discussed in sections 1.6 and 1.7.1. What figure 47 might demonstrate is that when the fast-ice contour is removed as a target, the previously described negative interactions between the fast-ice contour which is spatially different from the other predicted cumulative contours are not occurring. As a result, the reduced classes model

is assigning less importance to the area, compared to the baseline model in figure 46 where Frans Josef Land is assigned high importance as potentially the fast-ice cumulative contour is covering that area only. This might be a consequence of training the baseline model, since the error backpropagated through the network had an additional fast-ice component mainly situated around the land-masses in the scene, causing the model to be tuned towards lowering errors in that area. Consequently, when the baseline model is making predictions, the land-covered pixels are attenuated more since the model has been explicitly fed information of the land pixels through the fast-ice contour. On the other hand, the reduced classes model which have not targeted the fast-ice contour explicitly, have been trained on cumulative contours which resemble each other. In summary, the important pixels may to some degree resemble the shared sea ice contour areas in the cumulative contours, such that when a cumulative contour which offers a significantly limited spatial extent compared to the other contours is used as target, the important regions of the image is reduced to the shared regions between all contours. It is believed that this is a consequence of the shared decoder, due to the previously described negative interactions and effect on the loss each field has.

## 1.10 decreased skill from 2-meter temperature

It was shown in figure 41 that model performance stochastically improved when not including 2-meter temperature from AROME Arctic. A repeat run proved that this result was consistent (orange line figure 41). Hence, despite deeming AROME Arctic as a significant contributor to model performance, figure 41 indicate that this is due to the two wind components. Despite 2-meter temperature causing model performance to decrease, figure 42 and 43 show that the model is strongly fitted to the 2-meter temperature, especially during summer and autumn, since both modifications cause model performance to deteriorate. Since AROME Arctic in large is covered by sea ice (Müller et al., 2017), many of the atmospheric variables are directly affected by the sea ice representation (Müller et al., 2017; Batrak et al., 2018; Batrak and Müller, 2019; Müller et al., 2023). For 2-meter temperature, a rapid decrease in temperature occurs at the sea ice edge (Müller et al., 2023) (see also figure 51 (a) and 52 (a)). Thus, it is likely that the deep learning system is able to infer some correlation between the recent ice chart and 2-meter temperature field in terms of defining the predicted sea ice edge. Which might explain why the model is strongly fitted towards 2-meter temperature.

Figure 49 showed that the highlighted regions extent beyond the sea ice edge for a model without 2-meter temperature as a predictor, when compared to the baseline model showed in figure 48. Inspecting figure 49 while at the same time considering that the baseline model is potentially strongly fitted to the 2-meter temperature as discussed above, it may be the case that the model without 2-meter temperature is attempting to remedy missing

information related to the sea ice edge by looking beyond the sea ice edge. In some scenes, such as figure 49 (a) and (b), the model is also considering the land-covered pixels from mainland Norway and Russia as important information, which may be due to the land-sea mask predictor consistently providing the model with a signal covering that area. Furthermore, autumn (this also occurs to a degree during summer), the model without 2-meter temperature is able to deem pixels related to sea ice in the fjords of Svalbard as important, despite a 2-meter temperature contour following the sea ice edge as seen in figure 50 (d). Hence, it seems that the 2-meter temperature predictor implements a spatial inspection barrier determining if the model is using pixels outside the sea ice edge or not. This might occur due to the possible strong correlation inferred by the deep learning system between the recent sea ice chart and 2-meter temperature, such that when the two predictors are working in tangent they amplify the position of the sea ice edge. Consequently, this causes the model to distinguish between whether each pixel belongs inside the sea ice edge or not, with only pixels inside the inferred sea ice edge having importance when forecasting a scene. This can have a severe limiting effect on model performance, since information is restricted to pixels only inside the recent sea ice chart.

### 1.11 Interpreting the case study

A case study was conducted for the prediction from the test data which achieved the highest NIIEE, with the input, target and prediction presented in figure 51. Figure 51 (c) demonstrate the smoothing effect of the deep learning prediction as discussed in section 1.7.2. However, given the operational use case intended for the forecasts, the smoothness represents variability and to some degree uncertainty which is preferred over more detailed forecasts which are at a higher risk of wrongful predictions. From a user perspective, a generalized view of the sea ice development may prove more valuable than a detailed resolve of certain sea ice structures (Murphy, 1993).

In section 1.8.1 it was noted that the model was very strongly fitted to the recent ice chart, and as a result the output is expected to resemble an adjusted version of the input. This is shown in figure 51 through two examples. Firstly, notice the red contour directly north of the easternmost part of Frans Josef land in the predictor. In the target, this contour is seemingly cut off, whereas in the deep learning forecast, the contour itself has been deformed rather than removed. The second example are the gradually decreasing contours between Frans Josef Land and Novaya Semlya, which in the target has seen a rapid growth in the red contour nearly covering up all of the orange contour. For the deep learning forecast, the contours from the predictor are mostly preserved albeit smoothened.

Based on the strength of the 2-meter temperature in figure 52 (a), the temperatures above the sea ice edge seem vary around 273K, which might imply that the scene in figure

51 is mainly wind driven. However, it is not assumed that the deep learning system is interpreting the temperature values, since the convolutional layers derive features from gradients. With respect to the discussion in section 1.10, the spatial inspection barrier should follow a combination of the sea ice edge in 51 (a) and the temperature gradients which can be inferred visually in figure 52.

The atmospheric conditions are shown in figure 52. Firstly, southeast of Svalbard there appear to be winds blowing in the positive y-direction, as well as slightly negative with close to 0 m/s x-winds, for where the pixels are assumed to be important (see section 1.9). Thus the winds are blowing in a direction which appear to be parallel along the sea ice edge, directed towards the bottom of the scene. From the direction of the wind vector, Ekman transport is occurring directed away from the sea ice edge. Consequently, it is expected that due to the Ekman transport the sea ice edge will expand in the same direction, and it can be seen in figure 51 (c) that the deep learning system have inferred this response by predicting a wider sea ice edge in the area. Although the target sea ice chart have moved less, a similar response seems to have happened.

Finally, figure 53 give a spatial representation of where the forecast error is located. Based on the IIEE distribution, it is clear that the opening occurring in the sea ice east of svalbard has a large role in increasing the NIIEE. As there are no indication of a long term anomaly in the 2-meter temperature field figure 52 (a), it is expectedly difficult to predict that such an rapid sea ice loss event may happen. On the contrary, given the discussion in section 1.8.2, a warm anomaly located above the region could have prompted the model to decrease the sea ice concentration. Although it might have been shown in the linear sea ice trend, figure 41 and 43 demonstrated that the trend was not utilized by the model at all, thus it was not considered for the case study. If there was a hole present in the region in the predictor, it may be more likely that the model would deform that structure. Given that the rapid sea ice loss is an unlikely event, which is at most covered by a very limited number of samples and not generalizable to the extent as the sea ice edge, figure 51 demonstrates that spurious sea ice formation / decline events are not well resolved by the deep learning system.

## References

- Araujo, A., Norris, W., and Sim, J.: Computing Receptive Fields of Convolutional Neural Networks, *Distill*, 4, <https://doi.org/10.23915/distill.00021>, 2019.
- Batrak, Y. and Müller, M.: On the warm bias in atmospheric reanalyses induced by the missing snow over Arctic sea-ice, *Nature Communications*, 10, <https://doi.org/10.1038/s41467-019-11975-3>, 2019.
- Batrak, Y., Kourzeneva, E., and Homleid, M.: Implementation of a simple thermody-

- dynamic sea ice scheme, SICE version 1.0-38h1, within the ALADIN–HIRLAM numerical weather prediction system version 38h1, Geoscientific Model Development, 11, 3347–3368, <https://doi.org/10.5194/gmd-11-3347-2018>, 2018.
- Carrasco, A., Øyvind Saetra, Burud, A., Müller, M., and Melsom, A.: PRODUCT USER MANUAL For Arctic Ocean Wave Analysis and Forecasting Products ARCTIC\_ANALYSIS\_FORECAST\_WAV\_002\_014, Tech. rep., 2022.
- Crawshaw, M.: Multi-Task Learning with Deep Neural Networks: A Survey, <https://doi.org/10.48550/ARXIV.2009.09796>, 2020.
- DeVries, T. and Taylor, G. W.: Learning Confidence for Out-of-Distribution Detection in Neural Networks, <https://doi.org/10.48550/ARXIV.1802.04865>, 2018.
- Fritzner, S., Graversen, R., and Christensen, K. H.: Assessment of High-Resolution Dynamical and Machine Learning Models for Prediction of Sea Ice Concentration in a Regional Application, 125, <https://doi.org/10.1029/2020jc016277>, neural Networks for predicting Sea-Ice concentration are only slightly more accurate than persistence forecasting for short-term predictions., 2020.
- Fukushima, K.: Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, Biological Cybernetics, 36, 193–202, <https://doi.org/10.1007/bf00344251>, 1980.
- Goessling, H. F. and Jung, T.: A probabilistic verification score for contours: Methodology and application to Arctic ice-edge forecasts, Quarterly Journal of the Royal Meteorological Society, 144, 735–743, <https://doi.org/10.1002/qj.3242>, 2018.
- Goessling, H. F., Tietsche, S., Day, J. J., Hawkins, E., and Jung, T.: Predictability of the Arctic sea ice edge, Geophysical Research Letters, 43, 1642–1650, <https://doi.org/10.1002/2015gl067232>, 2016.
- Grigoryev, T., Verezhenskaya, P., Krinitskiy, M., Anikin, N., Gavrikov, A., Trofimov, I., Balabin, N., Shpilman, A., Eremchenko, A., Gulev, S., Burnaev, E., and Vanovski, V.: Data-Driven Short-Term Daily Operational Sea Ice Regional Forecasting, Remote Sensing, 14, <https://doi.org/10.3390/rs14225837>, URL <https://www.mdpi.com/2072-4292/14/22/5837>, 2022.
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, <https://doi.org/10.1002/qj.3803>, URL <https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803>, 2020.
- Hibler, W. D.: A Dynamic Thermodynamic Sea Ice Model, Journal of Physical Oceanogra-

- phy, 9, 815–846, [https://doi.org/10.1175/1520-0485\(1979\)009<0815:adtsim>2.0.co;2](https://doi.org/10.1175/1520-0485(1979)009<0815:adtsim>2.0.co;2), 1979.
- Jha, A., Kumar, A., Pande, S., Banerjee, B., and Chaudhuri, S.: MT-UNET: A Novel U-Net Based Multi-Task Architecture For Visual Scene Understanding, in: 2020 IEEE International Conference on Image Processing (ICIP), IEEE, <https://doi.org/10.1109/icip40778.2020.9190695>, 2020.
- Køltzow, M., Schyberg, H., Støylen, E., and Yang, X.: Value of the Copernicus Arctic Regional Reanalysis (CARRA) in representing near-surface temperature and wind speed in the north-east European Arctic, *Polar Research*, 41, <https://doi.org/10.33265/polar.v41.8002>, 2022.
- Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P.: Focal Loss for Dense Object Detection, 2017.
- Liu, G., Reda, F. A., Shih, K. J., Wang, T.-C., Tao, A., and Catanzaro, B.: Image Inpainting for Irregular Holes Using Partial Convolutions, <https://doi.org/10.48550/ARXIV.1804.07723>, 2018.
- Luo, W., Li, Y., Urtasun, R., and Zemel, R.: Understanding the Effective Receptive Field in Deep Convolutional Neural Networks, <https://doi.org/10.48550/ARXIV.1701.04128>, 2017.
- 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.
- Murphy, A. H.: What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting, *Weather and Forecasting*, 8, 281–293, [https://doi.org/10.1175/1520-0434\(1993\)008<0281:wiagfa>2.0.co;2](https://doi.org/10.1175/1520-0434(1993)008<0281:wiagfa>2.0.co;2), 1993.
- Müller, M., Batrak, Y., Kristiansen, J., Køltzow, M. A. Ø., Noer, G., and Korosov, A.: Characteristics of a Convective-Scale Weather Forecasting System for the European Arctic, *Monthly Weather Review*, 145, 4771–4787, <https://doi.org/10.1175/mwr-d-17-0194.1>, 2017.
- Müller, M., Batrak, Y., Dinessen, F., Grote, R., and Wang, K.: "Submitted" Challenges in the description of sea-ice for a kilometer-scale weather forecasting system, 2023.
- Nair, V. and Hinton, G.: Rectified Linear Units Improve Restricted Boltzmann Machines Vinod Nair, vol. 27, pp. 807–814, 2010.
- Palerme, C., Müller, M., and Melsom, A.: An Intercomparison of Verification Scores for Evaluating the Sea Ice Edge Position in Seasonal Forecasts, *Geophysical Research Letters*, 46, 4757–4763, <https://doi.org/10.1029/2019gl082482>, 2019.
- Ronneberger, O., Fischer, P., and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, in: *Lecture Notes in Computer Science*, pp. 234–241, Springer International Publishing, [https://doi.org/10.1007/978-3-319-24574-4\\_28](https://doi.org/10.1007/978-3-319-24574-4_28), 2015.
- 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 ensemble prediction model for the Barents Sea and Svalbard, p. 20, 2022.
- Sakov, P., Counillon, F., Bertino, L., Lisæter, K. A., Oke, P. R., and Korablev, A.: TOPAZ4: an ocean-sea ice data assimilation system for the North Atlantic and Arctic, *Ocean Science*, 8, 633–656, <https://doi.org/10.5194/os-8-633-2012>, 2012.
- Shorten, C. and Khoshgoftaar, T. M.: A survey on Image Data Augmentation for Deep Learning, *Journal of Big Data*, 6, <https://doi.org/10.1186/s40537-019-0197-0>, 2019.
- Spren, 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.
- Spren, G., Kwok, R., and Menemenlis, D.: Trends in Arctic sea ice drift and role of wind forcing: 1992-2009, *Geophysical Research Letters*, 38, n/a–n/a, <https://doi.org/10.1029/2011gl048970>, 2011.
- Tonboe, R., Lavelle, J., Pfeiffer, R.-H., and Howe, E.: Product User Manual for OSI SAF Global Sea Ice Concentration, Tech. Rep. 1.6, Danish Meteorological Institute, 2017.
- 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., 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.
- Vinogradova, K., Dibrov, A., and Myers, G.: Towards Interpretable Semantic Segmentation via Gradient-Weighted Class Activation Mapping (Student Abstract), *Proceedings of the AAAI Conference on Artificial Intelligence*, 34, 13 943–13 944, <https://doi.org/10.1609/aaai.v34i10.7244>, 2020.
- Wagner, P. M., Hughes, N., Bourbonnais, P., Stroeve, J., Rabenstein, L., Bhatt, U., Little, J., Wiggins, H., and Fleming, A.: Sea-ice information and forecast needs for industry maritime stakeholders, *Polar Geography*, 43, 160–187, <https://doi.org/10.1080/1088937x.2020.1766592>, 2020.
- Wang, L., Scott, K., and Clausi, D.: Sea Ice Concentration Estimation during Freeze-Up from SAR Imagery Using a Convolutional Neural Network, *Remote Sensing*, 9, 408, <https://doi.org/10.3390/rs9050408>, 2017.
- 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.
- Williams, T. D., Bennetts, L. G., Squire, V. A., Dumont, D., and Bertino, L.: Wave–ice interactions in the marginal ice zone. Part 1: Theoretical foundations, *Ocean Modelling*, 71, 81–91, <https://doi.org/10.1016/j.ocemod.2013.05.010>, 2013.
- Yu, X., Rinke, A., Dorn, W., Spren, G., Lüpkes, C., Sumata, H., and Gryanik, V. M.: Evaluation of Arctic sea ice drift and its dependency on near-surface wind and sea ice

- conditions in the coupled regional climate model HIRHAM–NAOSIM, *The Cryosphere*, 14, 1727–1746, <https://doi.org/10.5194/tc-14-1727-2020>, 2020.
- 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.
- Zhang, Z., Luo, P., Loy, C. C., and Tang, X.: Facial Landmark Detection by Deep Multi-task Learning, in: *Computer Vision – ECCV 2014*, pp. 94–108, Springer International Publishing, [https://doi.org/10.1007/978-3-319-10599-4\\_7](https://doi.org/10.1007/978-3-319-10599-4_7), 2014.