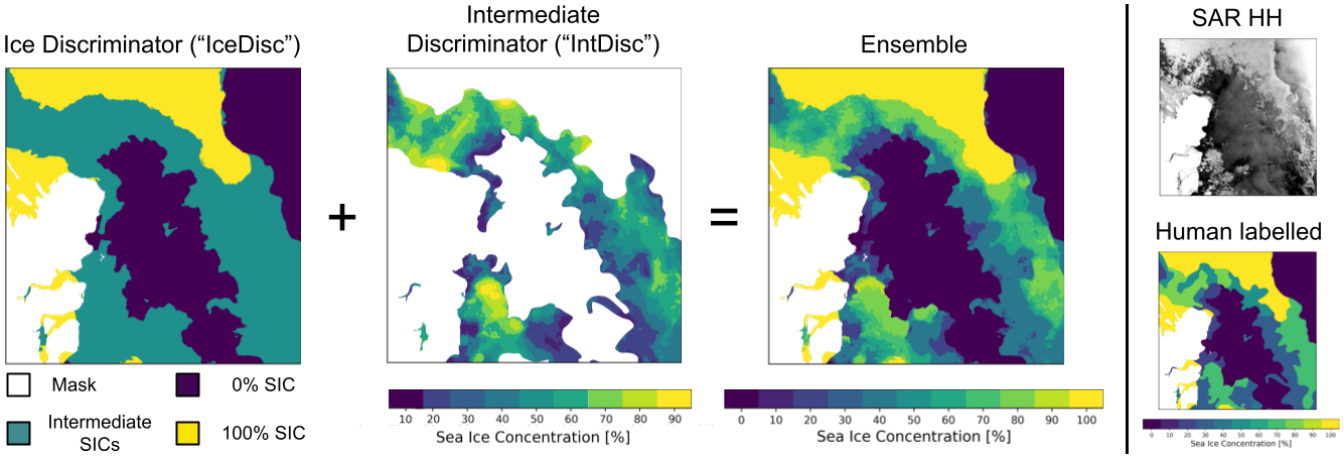
AI4SeaIce: Task Separation and Multistage Inference CNNs for Automatic Sea ice Concentration Charting

Andreas Stokholm[1, 2], Andrzej Kucik[2], Nicolas Longépé[2], and Sine Munk Hvidegaard[1]

[1]DTU Space, Department of Space Research and Technology, Technical University of Denmark (DTU), Elektrovej, building 327, Kgs. Lyngby, 2800, Denmark.
[2]ϕ-lab, European Space Research Institute (ESRIN), The European Space Agency (ESA), Largo Galileo Galilei, 1, Frascati RM, 00044, Italy.

Correspondence: Andreas Stokholm (stokholm@space.dtu.dk)



Abstract. We investigate how different Convolutional Neural Network (CNN) U-Net models specialised in addressing partial labelling tasks related to mapping Sea Ice Concentration (SIC) can improve performance. We use Sentinel-1 SAR images and human-labelled ice charts as the reference to train models that benefit from advantages gained from different model optimisation objectives by utilising a multistage inference scheme. We find our multistage model inference approach that ~~apply~~ applies a classification (CrossEntropy or Earth Mover's Distance squared) optimised model to separate open water, intermediate SIC and fully covered ice in conjunction with a regression (Mean Square Error or Binary CrossEntropy) optimised model, that assigns specific intermediate classes, to perform the best. To evaluate the models we introduce several specific metrics illustrating the performance in key areas, such as the separation of macro classes, intermediate class, and an accuracy metric better encapsulating uncertainties in the reference data. We achieve R^2-score of $\sim$93%, similar to state-of-the-art in the literature (Kucik and Stokholm, 2023). However, our models exhibit significantly better open water and 100% SIC ~~detections~~segmentations. The multistage approach synergises high open water and fully covered sea ice accuracies achieved with classification optimised objectives with good intermediate class performance obtained by regressional loss functions. In addition, our findings indicate that the number of classes that the intermediate concentrations are ~~compressed~~ merged into does not influence the result sig-

nificantly but rather it is the loss function used to optimise the model which assigns the specific intermediate class that has the largest impact.

1 Introduction

Charting the ever-changing sea ice is important for navigation in the remote and cold Arctic ~~to both circumnavigate and traverse sea ice~~ for traversing safely and quickly. ~~Effective navigation thus makes high-resolution~~ High-resolution sea ice charts detailing the local sea ice conditions ~~indispensable. This is particularly relevant for~~ are indispensable for effective navigation. Due to sparse infrastructure, local and indigenous populations ~~, e.g. around~~ along the Greenland coast ~~. Livelihoods in these regions depend on fishing and infrastructure is sparse, where goods and people are primarily transported by boat~~rely on goods transported by ship while many livelihoods depend on fishing, which makes ice charts important economic enablers. However, with the diminishing Arctic sea ice (Perovich et al., 2020), the Northern trade routes are also becoming increasingly relevant. This could offer alternative shorter shipping avenues connecting the Atlantic and the Pacific oceans through the Arctic, promising lucrative time and cost savings (Bekkers et al., 2017). Furthermore, decreasing sea ice cover is believed to result in more dynamic ice conditions (Boutin et al., 2020), which could continue to ~~cause~~ exacerbate hazardous conditions. For these reasons, the ability to map sea ice ~~could~~ should be considered a critical infrastructure component in the Arctic. ~~Furthermore, high-resolution~~ High-resolution sea ice information could also benefit weather and climate models by incorporating higher spatial details in the ice cover such as leads that allow for interactions between the ocean and atmosphere otherwise insulated by the sea ice.

1.1 Context

For the past 50 years, professional sea ice analysts at the Greenland Ice Service, a part of the Danish Meteorological Institute (DMI), and other similar institutions, e.g. the Canadian or Norwegian national ice services, have mapped the Arctic sea ice with a variety of methods ranging from airborne campaigns to satellite measurements. The Arctic is particularly challenging to monitor because of its ~~distant location, sparse infrastructure and vastness of the area~~vastness, remoteness and lack of infrastructure. Therefore, satellite observations offer appealing advantages with frequent revisit times and large coverage. However, conventional optical imagery is disproportionately affected by cloud cover, which has an ~~almost indistinguishable albedo reflection~~ albedo reflection almost indistinguishable from sea ice. In addition, the long Arctic polar night offers little to no sunlight for a significant portion of the year and incapacitates the use of optical imagery in this recurring period. Instead, ice analysts utilize microwave measurements from satellites. Here, passive microwave measurements, from e.g. the AMSR2 instrument onboard the JAXA GCOM-W1 satellite, are excellent for covering the whole Arctic. On the contrary, the instrument offers insufficient spatial detail for precise tactical navigation or lead detection with resolutions ranging from 35x62 km to 3x5 km per pixel, depending on the measurement frequency (6.925 - 89 GHz) (Kasahara et al., 2012). Consequently, Synthetic Aperture Radar (SAR) images are the backbone for sea ice charting ~~,~~ as they offer versatile measurements ~~in a high spatial resolution (10 - 40 m pixel spacing)~~ with pixel spacing typically between 10-40 m independent of sun illumination and clouds. However, passive

microwave, optical and thermal-infrared imagery are incorporated in the manual charting process when available and beneficial ((Saldo et al., 2021), manual). The main drawback of SAR is interpretability as radar backscatter is dependent on the surface properties and roughness, and not on the light emitted from an object. In addition, open-water and sea ice can be ambiguous in their electromagnetic texture appearance depending on e.g. weather conditions. Therefore, the charts are produced by a manual in-depth interpretation by experienced ice analysts and drawn using Geographical Information System (GIS) software. Naturally, this is ~~a~~ resource- and time-consuming, constraining the number of charts produced on a given day, by the amount of manpower and the data availability. This motivates the desire to fully or partially automate the charting process. Advantages include increasing the number of produced charts with shorter delays between image acquisition and product availability, shorter production duration, and the possibility of scaling the mapping coverage at little cost, e.g. to cover the entire Arctic.

1.2 Previous and other works

Automating sea ice charting has been studied for decades, and contemporary attempts have highlighted Convolutional Neural Networks (CNNs) as a strong contender to solve this challenge with image segmentation. Such approaches were first publicised by Wang et al. (2016) followed by additional entries in Wang et al. (2017a, b), which provided compelling proof of the validity of the approach to map the Sea Ice Concentration (SIC). However, in this early study, network complexity, data quantity and coverage was limiting factor. In 2020, the Automated Sea Ice Product (ASIP) project launched its first version of an open-source deep-learning dataset, the ASIP Sea Ice Dataset (ASID-v1) (Malmgren-Hansen et al., 2020). Initial results using the dataset were published in Malmgren-Hansen et al. (2021) with a custom-built CNN architecture using data fusion of SAR and AMSR2 and a regression-based optimisation approach. These early results were the first to apply large datasets of multiple 100 GBs for training and highlighted challenges with correctly classifying open water and ice in the Sentinel-1 SAR subswath transitions that contain noise and speckle (further introduced in 2.1). Heidler et al. (2021) was able to highlight that a larger receptive field could improve performance over Malmgren-Hansen et al. (2021). The ASIP project was continued partly as the European Space Agency's (ESA) project: AI4Arctic and produced the second version of the dataset, ASID-v2, in 2021 Saldo et al. (2021), as one of the ESA AI Ready Earth Observation (AIREO) datasets (information regarding the dataset is presented in Sec. 2). The dataset has been used to develop and publish new CNN-related works, such as Tamber et al. (2022) and the AI4SeaIce article series (Stokholm et al., 2022; Kucik and Stokholm, 2022, 2023). Other parallel efforts include the ExtremeEarth project described in Koubarakis et al. (2021) with its polar use-case with publications such as Khaleghian et al. (2021) focusing on the type of sea ice, which has also been investigated in a non-related study in Boulze et al. (2020). Other notable ice mapping entries in the literature include Radhakrishnan et al. (2021) applying curriculum learning. And finally de Gelis et al. (2021), which improved previous results and underlined challenges in correctly classifying areas with ambiguous SAR signatures. Approaches to overcoming these challenges were examined in Stokholm et al. (2022), investigating the effects of increasing the receptive field of the CNN models for various ambiguous SAR textures. The newest study on the topic (at the time of writing) is Kucik and Stokholm (2023), which investigates different loss functions for CNN model optimisation.

This paper is the 4th entry into the series; AI4SeaIce, investigating applying AI to automatically map sea ice based on SAR imagery (Stokholm et al., 2022; Kucik and Stokholm, 2022, 2023).

1.3 Objective

Ice analysts are capable of charting a variety of information regarding sea ice. The two most important aspects are the amount of sea ice in an area and the type of sea ice, which is a proxy for its thickness. The former is the focus of this study and is described as the amount of sea ice ~~in relation~~ relative to open water and is denoted as the SIC. SIC is defined by the World Meteorological Organisation (WMO) as part of the ~~SIGRID-3 code~~Sea Ice GeoReferenced Information and Data code, also known as SIGRID-3. In this study, the SIC is given as a percentage from 0-100% in discrete increments of 10%, i.e. 11 ice concentration classes. The increments exhibit substantial relationships between each class, i.e. 50% is more approximate to 60% than e.g. 40% or 70%, which will be referred to as *interclass* relationships in the proceeding. The charting process involves identifying areas of sea ice in SAR images and drawing polygons of comparable and relatively homogenous sea ice, which are assigned ice parameters, such as the SIC.

The supervised machine-learning segmentation task may seem straightforward as a regression problem, however, a previous AI4SeaIce study in Kucik and Stokholm (2023) indicate that this may not be a prevalent strategy. The study carried out a comparison of analogies of how to interpret the machine-learning task at hand and the effects of different optimisation objectives; regression-based Mean Square Error (MSE) and Binary CrossEntropy (BCE), and classification-based categorical CrossEntropy (CE) and the squared Earth Mover's Distance (EMD^2). A key takeaway of the study was the observation that models optimised with classification objectives were significantly better at correctly predicting open-water (0% SIC) and sea ice polygons of 100% sea ice. On the contrary, the regression objectives produced models superior at assigning the intermediate (10-90%) SIC, i.e. the classes with the strongest interclass relationships. While the regression-based models achieved higher scores with respect to numerical similarity, the models produced charts exhibiting high-frequency class transitions not present in the human-labelled charts, which could be interpreted as a higher resolution but it could also be a potential downside, depending on the ice chart and user preferences.

A natural question arises from the Kucik and Stokholm (2023) study — could we combine the advantages from the classification and regression optimised models? As there is a clear discrepancy in the performance related to the optimisation objective, we suggest an approach of dissecting the overall SIC assignment problem into two smaller ones, handled by different specialised models — in effect creating a multistage model inference.

Therefore, we investigate utilising a model to discriminate whether pixels are open-water, an (any) intermediate class of SIC or fully covered sea ice (the IceDisc model) to in effect separate the largest (macro) classes. This is combined with another model to further identify the intermediate SIC (the IntDisc model). The approach allows the models to be optimised with different loss objectives, but potentially different architectures could be used. However, questions related to how we ~~compress~~ merge the intermediate classes are imminent. Is one class sufficient to encapsulate 9 classes from 10 to 90%, or perhaps no ~~compression~~ merging is needed at all? With such large diversity, we propose to include an examination, of whether the number of classes influences the model's ability to separate the macro classes. This aspect also ties nicely into the debate between members within the sea ice community and among AI practitioners with regards to whether 11 classes are simply too many

classes considering the uncertainty associated with the labelling procedure (see Subsec. 2.2). Here, we can provide empirical evidence of whether the number of intermediate SIC classes impacts the model's ability to separate the macro classes.

For completeness, we expand the experiment to also include an IceDisc, which only differentiates between open water and 100% fully covered sea ice. This is coupled with an IntDisc model that assigns both intermediate and 100% SIC. To provide a perspective of the performance, we naturally compare these approaches with similarly trained models presented in Stokholm et al. (2022); Kucik and Stokholm (2023) with both classes un- and weighted loss functions.

2 Data - the ASID-v2 dataset

The research is conducted utilising the ASID-v2 (Saldo et al., 2021), compiled by DMI, the Technical University of Denmark (DTU), and Nansen Environmental and Remote Sensing Center (NERSC) and released on October 2, 2020. A total of 452 co-located and georeferenced scenes acquired between March 14 2018 and May 25 2019 are included and distributed across the Greenland coast with the majority from the mid-East, South and mid-West and sparse appearance in the North. Among others, the dataset consists of professionally drawn sea ice charts, which we treat as reference data, and Sentinel-1 dual polarised HH and HV SAR images. We limit our investigation to utilising these data types to further our understanding of how well standalone Sentinel-1 SAR-trained CNN models can be used to replicate human-labelled ice charts, despite not having all data sources available.

In Fig. 1 an example of an HH and HV polarised SAR scene with the corresponding professionally produced SIC chart is illustrated. The scene is from Northeast Greenland and covers $\sim 400km^2$. Here, land in the lower-left region of the image is masked and illustrated as white pixels. Sea ice generally appears brighter than still or calm open water in both HH and HV polarised SAR images, which can be ~~contributed~~ attributed to the relatively rougher surface in relation to the C-band electromagnetic radar wave that scatter back more of the incoming radar wave. Exceptions to this in the HH channel can occur during dynamic waters excited by strong winds, inducing stronger backscatter values that can be higher than sea ice. In addition, the incidence angle dependency is more prominent for open water, which can also produce backscatter values higher than sea ice.

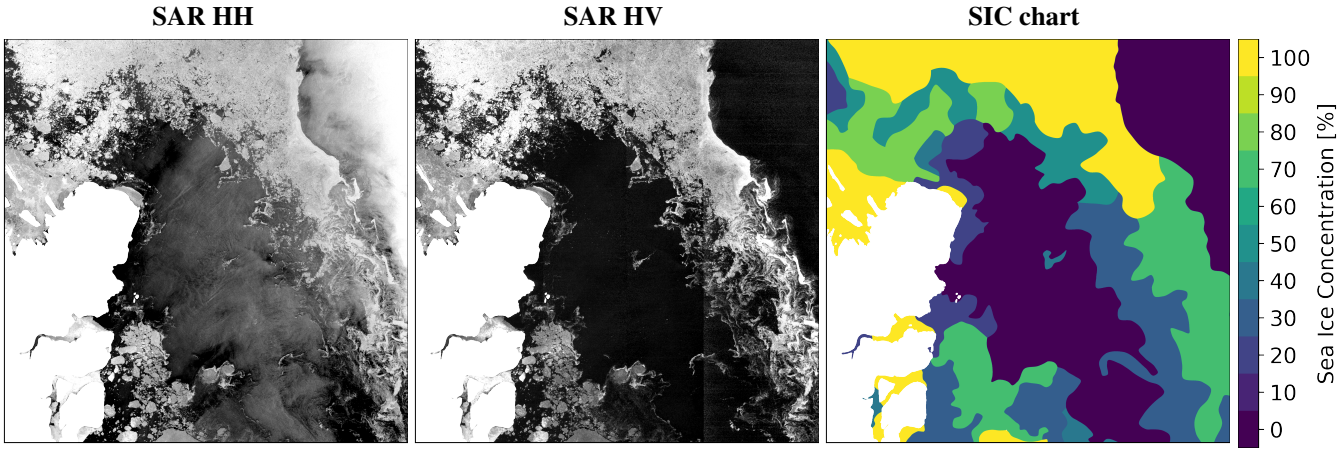


Figure 1. Sample scene – Fram Strait, Northeast Greenland. The scene was acquired on August 22, 2018. **a-b**) HH and HV SAR images, respectively, with values, clipped at the 5 and 95% percentiles. **c**) corresponding human-drawn SIC chart (reference).

The disparity between sea ice and open water is largest in the HV image with a narrower dB dynamic range de Gelis et al. (2021), though the strongest backscatter signal is in the HH image. Investigating the SAR image in Fig. 1, we see sea ice extending from the North towards the South surrounding an area of open water. In the far right portion of the SAR image, we see a region of open water, which appears very bright. Such a phenomenon is caused by a relatively narrow incidence angle of the SAR for the observational geometry. Examining the SIC chart, we see many polygons highlighting regions of varying SICs.

2.1 SAR

The Sentinel-1A and B satellites were formerly operated as a constellation with Sentinel-1B unavailable from December 2021. The ASID-v2 dataset was compiled before the malfunction, and thus we utilize data from both satellites. The satellites are instrumented with a C-band SAR operating at 5.410 GHz frequency or a wavelength of 5.5 cm (Torres et al., 2012). The utilised data is the level 1 Ground Range Detected Medium resolution product, measured in the Extra-Wide operational mode with a native pixel spacing of 40 m and a radar resolution of 93×87 m (range $\times$ azimuth). In the ASID-v2 dataset, there are two different SAR noise corrections available; the ESA Instrument Processing Facilities (IPF) v2.9 and the NERSC noise correction (Park et al., 2018, 2019). In correspondence with the findings from Stokholm et al. (2022), we utilise the NERSC noise correction. At the time of writing, both noise corrections have since been updated; the current ESA IPF version is 3.61 and the newest NERSC version is described in Korosov et al. (2022).

2.2 Sea ice charts

Sea ice charts are based on professional interpretations of SAR images and represent snapshots of the ice condition at the capture time. The ice information is conveyed using the 'Egg Code', which follows the ~~World Meteorological Organization (WMO) code; Sea Ice GeoReferenced Information and Data (~~WMO code; SIGRID3~~)~~, and is represented graphically as poly-

gons of fairly homogeneous areas of sea ice. The process is steered by common guidelines but is fundamentally a creative and individual interpretation. Studies have suggested that ice analysts assign concentrations that can vary on average by 20% and in worst cases up to 60% (Karvonen et al., 2015), which disproportionately affect the intermediate SIC classes. Cheng et al. (2020) also document that low SICs (10-30%) are overestimated while middle SIC classes (50-60%) can exhibit high variability with a wide spread. Regions, such as the edge of the sea ice cover, which have the potential for high maritime activity, often receive more attention from analysts. Despite these uncertainties, we treat the human-labelled SIC labels as reference data or ground truth and each pixel is treated as equally valid.

In the ASID-v2 dataset, ice charts are stored as polygons containing IDs for an associated look-up table containing the sea ice information for the specific polygon. Originally, 14 codes exist for SIC, which we compress into 11 classes from 0-100% in discrete increments of 10%, i.e. class 0 — class 10.

2.3 Data preparation

We further process the data using the same setup as in Stokholm et al. (2022). SAR and SIC images are downsampled from 40 m to 80 m pixel spacing by applying a 2×2 averaging kernel and a 2×2 max kernel, respectively. Pixels with land or no information are masked and aligned across the SAR and SIC, and rows and columns only containing masked pixels are discarded. Masked pixels are given values 0 in the SAR data and a new class 11 is established for the SIC, which is discounted during model training. Class 11 is not mentioned explicitly in the proceeding class bundling. The SAR data is normalized to the [-1, 1] range using maximum and minimum values of the data distribution.

Training batches of 32 patches are prepared with each patch containing 768^2 pixels, which are sampled across the training scenes based on the number of available pixels in each scene, i.e. scenes with more pixels will be sampled more often than those with fewer, as described in Stokholm et al. (2022). Batches are given with a random set of augmentations — one for each patch — and identical across SAR and SIC. We utilise the dihedral group: 0, 90, 180 or 270-degree rotations, as well as horizontal, vertical and two diagonal flips, i.e. 8 in total. Further, we give each patch a 50% chance of applying between 1-4 affine transforms with a random bounded magnitude; [-44.99, 44.99] degrees of rotation, $\pm$30% scaling, $\pm$30% translation, and $\pm$10 degrees of shearing.

2.4 Data split and distribution

From the 452 scenes in the ASID-v2 dataset, we select 306 for training and 23 for testing while discarding the remaining 123 scenes based on containing mainly open water while a few contain errors. Each scene with up to $\sim 5,000^2$ pixels, constituting roughly a train and test split of 9:1 in terms of pixel count, retaining the SIC class distribution. This distribution is skewed towards class 0 (open-water), 10 (100% sea ice) and 11 (masked pixels) being represented the most as highlighted in Stokholm et al. (2022). The remaining intermediate classes are relatively equally distributed. The test set was selected in collaboration with DMI, and scenes were selected to be particularly difficult by ice analysts to challenge trained models in tough conditions. The test set mirrors the training class distribution with marginally more intermediate-class pixels. We ensure that no cross-sampling occurs between the training and test scenes to prevent regional biases. However, some marginal leakage

may occur between scenes sampled consecutively or between acquisitions from satellite orbits where little sea ice drift has occurred.

3 Experimental setup

We train models with the U-Net CNN architecture (Ronneberger et al., 2015) containing 8 encoder-decoder blocks (16 and 32 filters in the first two and 64 filters in the remaining) for 100 epochs, each with 500 batches (training steps), similar to those utilised in Kucik and Stokholm (2023); Stokholm et al. (2022). We utilise the Adam optimiser with a fixed learning rate of 0.0001 and default PyTorch hyperparameters. As recommended in Huang et al. (2018), testing and inference are performed on entire unaugmented scenes. Experiments were performed on two Nvidia TeslaV100 SXM2 32GB GPUs using PyTorch version 1.8 and cuda 11.6.

Models are optimised with the CE and EMD^2 for the IceDisc models, and for the IntDisc models two alternative loss functions are included; MSE and BCE. Pixels containing masked pixels are discounted in both the loss calculation by multiplying relevant pixel loss values with 0 and during metric computation. In ~~the case of CE, the *ignore_index* argument is used instead, having the same effect. In~~ addition, we scale the loss of each patch in the batch by the ratio of valid pixels to masked pixels, giving larger weight to patches with more valid pixels. ~~More details are available, including a definition of the EMD^2 loss function in Kucik and Stokholm (2023)~~In Kucik and Stokholm (2023), a detailed analysis of various loss functions for the sea ice concentration classification was carried out. In this paper, we perform further experiments to explore further the potential of the loss functions: CE, BCE, MSE, and EMD2 discussed therein.

At the end of every epoch, models are tested and model parameters are stored. The epoch model parameters, which score the highest on the test set according to the chosen metrics described in subsections 3.1 and 3.2 are selected as the final model for that training cycle. To minimize the impact of random model initialisation conditions, we carry out all model experiments 3 times for each class configuration and every loss function. The best model for each class configuration and loss function is then further selected as a fair representation of a model trained with this approach could accomplish. This is also applied to the trained reference models. In total, 78 models are trained. We choose to show only the best-performing models for simplicity. Furthermore, the combination of the best-performing IceDisc and IntDisc models, and the regularly trained reference models give rise to 60 final model results.

3.1 Training IceDisc

We investigate how an IceDisc model can best discriminate between open-water, intermediate SICs and 100% sea ice by reducing the total number of classes by merging the intermediate ones. An overview of the class combinations is available in Tab. 1. In total, we attempt training models capable of predicting 6 different combinations of classes from 2 to 7 classes in addition to the regular 11 class models. In the first attempt, all sea ice classes are bundled together to create 2 classes - water (0) and ice (1). Afterwards, all intermediate classes are combined to create 3 classes - water (0), intermediate (1) and 100% (2). The number of classes is then gradually increased by 1 to dilute the number of intermediate SICs in a single class. For the

two-class configurations (6 and 7 classes) that are less intuitive; the models trained with 6 classes separate "little" sea ice (class 1) and a lot of sea ice (class 4 or 90%) more distinctively than the middle classes. A standalone class of 90% for the 6 and 7 class configurations were chosen because it is somewhat overrepresented compared to the other intermediate classes but less than 100%. IceDisc models are trained with classification objectives, i.e. CE and EMD^2.

To evaluate the IceDisc models, we define a new ~~metric that we denote~~ SIC-specific metric, *"MacroBins"*, ~~which is~~ an altered version of the Mean Producer Accuracy or the Average Class Accuracy. The ~~objective of this metric is to measure how well the model is capable~~ metric measures the models' capability of separating open water, intermediate classes and 100% sea ice. We define this metric as:

$$\text{MacroBins} = \frac{1}{3}\left(\frac{\text{water}_{TP}}{\text{water}_{NP}} + \frac{\text{int}_{TP}}{\text{int}_{NP}} + \frac{100_{TP}}{100_{NP}}\right) \tag{1}$$

where TP is the number of True Positives, NP is the Number of Pixels belonging to the class and *int* are intermediate class pixels. ~~In essence, this~~ Effectively, it is the average accuracy of water, 100% sea ice and whether true intermediate class pixels are predicted as any intermediate class. We will onwards refer to the latter as *int in int%*. As it is an average accuracy, the metric can illustrate the model's ability to separate the three macro classes in percentage. The MacroBins metric will be the main evaluation metric for the IceDisc models. In the case of binary 2-class models, MacroBins are replaced by the overall accuracy of the model.

3.2 Training IntDisc

IntDisc models are separated into two distinct discriminators — the ordinary IntDisc, which discriminates between intermediate classes (10-90%), and the Int100Disc, which differentiates between all sea ice classes (10-100%). The latter is designed to be applied with the IceDisc 2-class model. Training the IntDisc models is straightforward. We mask any pixel belonging to either the open-water or 100% sea ice classes enabling training on purely intermediate sea ice classes. Due to the strong interclass relationships, the IntDisc models are primarily evaluated on the R^2-score. The R^2-metric is also known as the Coef-

Table 1. Class (cls) combinations for the Ice Discriminator ("IceDisc"). The original 11 Sea Ice Concentrations are compressed to between 2-7 classes. Classes represent the percentage ranges expressed in the columns for each class configuration.

IceDisc	cls 0	cls 1	cls 2	cls 3	cls 4	cls 5	cls 6
2 cls	0%	10-100%					
3 cls	0%	10-90%	100%				
4 cls	0%	10-40%	50-90%	100%			
5 cls	0%	10-30%	40-60%	70-90%	100%		
6 cls	0%	10-20%	30-50%	60-80%	90%	100%	
7 cls	0%	10-20%	30-40%	50-60%	70-80%	90%	100%

Table 2. Overview of the Flexible Accuracy metric and which SICs are evaluated as correct.

SIC	0%	10%	20%	30%	40%	
Correct	0%	10-20%	10-30%	20-40%	30-50%	
SIC	50%	60%	70%	80%	90%	100%
Correct	40-60%	50-70%	60-80%	70-90%	80-90%	100%

ficient of Determination. It is often better in capturing the continuity of the distance between predictions and the ground truth. Accuracy and the Average Class Accuracy (ACA) are also reported. IntDisc models are trained with 4 different loss functions, 2 classification and 2 regression-oriented approaches, CE and EMD^2, and MSE and BCE, respectively. The Int100Disc is trained by masking pixels belonging to the open water class enabling discrimination on only ice classes.

3.3 Multistage Inference

Inference for the model combinations is performed by first producing the output maps for the IceDisc model on an entire test scene. If the class configuration contains more than 3 classes, all intermediate classes are compressed into a single class. Afterwards, the IntDisc performs inference on the same scene but utilises only the pixels determined by the IceDisc model to contain intermediate classes. Hereby, the two model outputs are combined into a single image. In the case of the IceDisc models with only 2 classes, the Int100Disc models produce outputs based on all the ice predictions.

3.4 Final evaluation

The multistage inferences are evaluated primarily on the R^2-score. However, a number of sub-metrics are also utilised, which show specific performance in key areas, such as the R^2-score measured on just the intermediate pixels (Int R^2). Furthermore, we define another metric that we denote "Flexible Accuracy" purposed to measure accuracy while accounting for some uncertainty in the ice charts. As highlighted in Tab. 2, intermediate pixels are evaluated as being correct if the model has produced a pixel belonging to a neighbouring class, e.g. if the correct SIC is 30%, 20 and 40% will also be deemed as correct. However, 0% and 100% SIC classes are evaluated as usual accuracy while 10% and 90% have extended ranges to 20% and 80%, respectively. Furthermore, the intermediate flexible-class accuracy is added showcasing flexible-class accuracy exempting accuracies for 0% and 100%.

To further analyse the performance of the multistage inference, we compare them with regularly trained models identical to Kucik and Stokholm (2023). These models are identical to the 11 class models used as IceDisc but the model parameters are instead selected based on the R^2-score. We choose to compare with the CE and EMD^2 loss functions in both an unweighted and weighted setting where the latter are trained using the frequency of classes as weights in the loss function, which is further elaborated in Kucik and Stokholm (2023).

4 Results

Initially, IceDisc models are investigated followed by the IntDisc models and evaluated quantitatively. Afterwards, the multistage inference combinations are examined quantitatively by applying the metrics characterised in subsection 3.4. Finally, a subset of multistage inference maps is highlighted and analysed qualitatively.

4.1 IceDisc results

The highest achieving IceDisc models for both the CE and EMD^2 losses are shown in Tab. 3 showcasing the MacroBins metric performance, as defined in subsection 3.1, and the open water and fully covered sea ice accuracy performance as well as the number of produced intermediate pixels belonging to intermediate classes. A row for each loss function with the average metrics is also included to provide a simple overview. In addition, two 11-class IceDisc models for each loss function are presented as references, similar to those trained in Kucik and Stokholm (2023). One is used as an IceDisc model and selected based on MacroBins while the reference 11-class models are chosen with respect to the R^2-score. The highest performances of the IceDisc models are highlighted in bold and underlining represent the best performance across the IceDisc loss functions.

Highest scores

Generally, we see that both CE and EMD^2 across all class combinations are quite capable of separating the three macro classes. Furthermore, both CE and EMD^2 optimised models score highest with 7 classes in terms of the MacroBins score, though the 3 class EMD^2 scored the second highest. In comparison, the 11-class IceDisc for both loss functions achieves the highest open-water accuracy.

2-class IceDisc

Inspecting the 2-class models, the performance for 0% is on par with most of the IceDisc models with additional classes but lower than the reference unweighted 11-class models. Otherwise, both 2-class models achieve high accuracies on the ice class.

CE IceDisc

For the CE IceDisc models, the 7-class model ~~outperforms the models~~ exceed those trained on fewer classes, particularly in terms of *int in int* and 100% sea ice while performing the worst at open water. Overall, the performances in the 0% category are very similar across the CE IceDisc models while there are fluctuations in the *int in int* and 100% categories. We also see that IceDisc models with 3-7 classes score ~~much~~ 4-7% better in terms of *int in int* compared to the 11-class IceDisc, which scores highest in 0% and 100% SIC.

EMD^2 IceDisc

With respect to the EMD^2 IceDisc models, we see fluctuating performances with the highest ~~macro classes distributed between different IceDiscs~~open water, *int in int* and 100% accuracies distributed among models with different numbers of classes. The 6-class model achieves the highest 100% accuracy but at the expense of the lowest *int int int* and 0% ~~and *int int int*~~ scores across all the CE and EMD^2 IceDisc models. The 3-class model on the other hand achieves the highest *int in int*-score. Again, the 11-class IceDisc model scores well on open-water but on par with respect to both *int in int* and 100% SIC.

CE vs EMD^2 IceDisc

The CE IceDisc models appear to generally score higher in terms of the *int in int* metric but lower in 0% and 100% compared with the EMD^2 IceDisc models as illustrated by the average metric scores.

IceDisc comparison to 11-class reference models

In comparison with the reference 11-class models, the unweighted optimised model scores high on both 0% SIC but underperforms in terms of *int in int*. The EMD^2 reference also scores high on 100%. In contrast, the 7-class CE model ~~approaches the~~

~~class weighted reference model on~~ exceeds both 11-class CE models by 6-8% on the *int in int* accuracy. The weighted models on the other hand scores abysmally on open water and fully covered sea ice.

Table 3. MacroBins results for the Ice Discriminators ("IceDisc") and the reference models. The best results for each bin are highlighted in bold separately for both loss functions, and the highest scores are underlined. In addition, mean metric scores are provided for the 3-11 class combinations. *Average of the 0%, *int in int*% and 100% performances; *int in int*% is a measure of the percentage of true intermediate pixels predicted as such. **Class weighted loss optimised model.

CE	Accuracy	0%	-	ice
2 cls	96.58%	95.60%	-	97.42%
	MacroBins*	0%	int in int%	100%
3 cls	90.55%	95.01%	83.91%	92.73%
4 cls	90.57%	95.25%	85.80%	90.65%
5 cls	90.46%	95.90%	85.39%	90.08%
6 cls	90.44%	95.35%	83.05%	92.93%
7 cls	**90.87%**	93.23%	**86.16%**	93.22%
11 cls	90.14%	**96.81%**	79.48%	**94.15%**
mean	**90.505%**	95.26%	**83.96%**	92.29%
Reference: CE				
11 cls	88.40%	97.76%	78.28%	89.15%
w** 11 cls	80.63%	81.58%	97.18%	63.11%

EMD^2	Accuracy	0%	-	ice
2 cls	96.60%	96.34%	-	96.83%
	MacroBins*	0%	int in int%	100%
3 cls	90.89%	95.29%	**85.43%**	91.96%
4 cls	90.72%	**95.88%**	82.36%	93.90%
5 cls	90.42%	95.76%	84.53%	90.98%
6 cls	89.46%	93.46%	80.36%	**94.55%**
7 cls	**90.90%**	94.72%	84.75%	93.23%
11 cls	90.60%	**97.27%**	81.76%	92.75%
mean	90.497%	**95.40%**	83.20%	**92.89%**
Reference: EMD^2				
11 cls	89.43%	97.86%	76.11%	94.33%
w** 11 cls	82.99%	76.40%	92.24%	80.33%

4.2 IntDisc results

The results for the IntDisc and Int100Disc models are presented in Tab. 4 highlighting performance on the R^2-metric, accuracy and the Average Class Accuracy (ACA). Again, the highest scores are highlighted in bold. For the IntDisc models, we see that R^2-scores are highest for MSE and BCE, whereas accuracy is highest for CE and EMD^2, though poor in comparison with the IceDisc ability to separate the macro classes. These results are in line with those presented in Kucik and Stokholm (2023) with respect to models trained with regression and classification loss functions.

For the Int100Disc models, the pattern is similar to the IntDisc models, where MSE and BCE achieve the highest R^2-scores but are inferior in terms of accuracy compared to CE and EMD^2. ~~The~~ However, the performance difference between the ~~loss functions within regression and classification is , however, performing very similarly~~classification loss functions across the 3 metrics is small and similar for the MSE and BCE-optimised models. The accuracy is a lot higher compared with the IntDisc models though expected since the much easier-to-predict 100% macro class is included.

Table 4. R^2, accuracy and **Average Class Accuracy (ACA) scores for the Intermediate Discriminators ("IntDisc") and Intermediate + 100% Discriminators ("Int100Disc"). The highest performance is marked in bold for the IntDisc and Int100Disc models separately. *Intermediate Sea Ice Concentration, **Average Class Accuracy.

Only Int*	R^2	Accuracy	ACA**	Int* + 100%	R^2	Accuracy	ACA*
CE	55.90%	35.64%	30.70%	CE	74.10%	54.17%	**31.91%**
EMD^2	59.55%	**36.23%**	**33.96%**	EMD^2	75.02%	**55.54%**	31.57%
MSE	**67.31%**	28.76%	24.43%	MSE	**80.77%**	49.66%	27.09%
BCE	66.52%	29.19%	26.47%	BCE	80.25%	49.40%	30.29%

4.3 Quantitative multistage inference results

The combined model inference results are displayed in Tab. 5 showcasing performance measured on the R^2-score, the intermediate class R^2-score, ~~multi-class~~ flexible-class accuracy, intermediate ~~multiclass~~ flexible-class accuracy as well as the MacroBins performance, which is repeated from Tab. 3 for convenience. Notice that the MacroBins are identical for each subtable of IceDisc and IntDisc combinations. To provide an overview, the mean scores have been added for the IceDisc/IntDisc combinations while omitting the mean MacroBins as these would be identical — note that these means do not include the 2-class information. Bold numbers represent the highest performance within IceDisc and IntDisc combinations, while underlined scores highlight the best metric performance across the IceDisc types. Red boxes indicate the models that are used in the qualitative analysis.

Highest scores

The highest R^2 and int R^2 scores are achieved by the EMD^2 ~~6-class~~ 3-class IceDisc and MSE IntDisc combination with 93.01%. In the same inference combination the 6-class version scores best on the ~~multi-accuracy metricwhile the topmost int multi-accuracy score~~ flexible-class accuracy metric. The topmost int flexible-class accuracy is obtained by the CE 7-class IceDisc and MSE IntDisc combination.

CE IceDisc

For the CE IceDisc-based models, we see a clear tendency for top-scoring metric performances for models trained with more classes. On the R^2-metric, the 6-class CE IceDisc-based models achieve the highest scores across all the IntDisc versions. The lowest mean scores are achieved by the CE IceDisc/IntDisc combinations while the best scores are achieved by regression-based MSE IntDiscs, though closely followed by the BCE IntDisc combinations.

Table 5. R^2, Intermediate (Int) R^2, Flexible Accuracy (FAcc), Int FAcc and MacroBins, and the average combination results for the multi-stage inferences. The highest scores are in bold for each combination. The best metric performance across combinations is underlined.

IceDisc: CE

classes	R^2	Int R^2	FAcc	Int FAcc	MacroBins
IntDisc: CE					
2	89.43%	43.17%	81.43%	52.40%	79.01%
3	90.15%	44.30%	81.99%	53.69%	90.55%
4	90.46%	**47.17%**	82.13%	55.41%	90.57%
5	90.08%	46.10%	82.34%	55.54%	90.46%
6	**90.82%**	45.48%	82.10%	53.37%	90.44%
7	90.16%	46.55%	81.97%	**55.92%**	**90.87%**
11	89.53%	35.97%	**82.35%**	51.01%	90.14%
mean	90.20%	44.26%	82.15%	54.16%	
IntDisc: EMD^2					
2	90.09%	43.05%	82.95%	54.86%	84.24%
3	90.69%	48.05%	82.92%	56.75%	90.55%
4	91.03%	**50.85%**	83.01%	58.32%	90.57%
5	90.47%	49.51%	83.19%	58.37%	90.46%
6	**91.32%**	49.16%	82.97%	56.27%	90.44%
7	90.95%	50.35%	82.94%	**59.14%**	**90.87%**
11	90.20%	39.99%	**83.27%**	54.05%	90.14%
mean	90.77%	47.99%	83.05%	57.15%	
IntDisc: MSE					
2	91.19%	55.44%	80.88%	57.12%	80.29%
3	91.40%	55.15%	83.29%	57.98%	90.55%
4	91.72%	**57.81%**	83.32%	59.35%	90.57%
5	91.13%	56.21%	83.44%	59.21%	90.46%
6	**91.91%**	55.82%	83.26%	57.22%	90.44%
7	91.49%	57.01%	83.23%	**60.11%**	**90.87%**
11	90.87%	46.12%	**83.68%**	55.42%	90.14%
mean	**91.42%**	**54.69%**	**83.371%**	**58.21%**	
IntDisc: BCE					
2	91.57%	56.73%	79.93%	59.21%	77.52%
3	91.16%	54.43%	83.18%	57.62%	90.55%
4	91.57%	**57.04%**	83.27%	59.16%	90.57%
5	90.93%	55.63%	83.41%	59.10%	90.46%
6	**91.74%**	55.09%	83.21%	57.04%	90.44%
7	91.37%	56.30%	83.14%	**59.83%**	**90.87%**
11	90.66%	45.38%	**83.61%**	55.20%	90.14%
mean	91.24%	53.98%	83.30%	57.99%	
Reference CE models					
11	90.79%	37.39%	81.96%	52.19%	88.40%
11 w	87.63%	52.42%	71.42%	62.40%	80.63%

IceDisc: EMD^2

classes	R^2	Int R^2	FAcc	Int FAcc	MacroBins
IntDisc: CE					
2	90.28%	43.13%	81.55%	51.66%	78.68%
3	91.75%	**47.56%**	82.35%	53.49%	90.89%
4	90.81%	40.79%	82.48%	50.64%	90.72%
5	91.32%	46.80%	82.18%	54.52%	90.42%
6	90.75%	46.49%	80.88%	50.93%	89.46%
7	90.85%	46.94%	82.29%	**54.71%**	**90.90%**
11	**92.04%**	46.51%	**82.62%**	52.29%	90.60%
mean	91.25%	45.85%	82.13%	52.76%	
IntDisc: EMD^2					
2	91.11%	44.26%	83.10%	54.15%	83.91%
3	**92.26%**	**51.27%**	83.27%	**58.09%**	90.89%
4	91.35%	44.24%	83.35%	55.92%	90.72%
5	91.85%	50.59%	83.04%	57.39%	90.42%
6	91.40%	50.64%	81.83%	54.10%	89.46%
7	91.38%	50.46%	83.17%	57.65%	**90.90%**
11	92.24%	49.74%	**83.43%**	54.99%	90.60%
mean	91.75%	49.49%	83.02%	56.36%	
IntDisc: MSE					
2	91.99%	55.57%	81.07%	56.60%	80.06%
3	**93.01%**	**58.49%**	83.57%	**59.09%**	90.89%
4	92.07%	51.22%	83.66%	56.95%	90.72%
5	92.46%	57.18%	83.28%	58.18%	90.42%
6	92.02%	57.37%	82.35%	55.80%	89.46%
7	92.12%	57.64%	83.55%	58.89%	**90.90%**
11	92.98%	56.53%	**83.84%**	56.31%	90.60%
mean	**92.44%**	**56.40%**	**83.374%**	**57.54%**	
IntDisc: BCE					
2	91.96%	56.84%	80.20%	**58.80%**	77.33%
3	92.72%	**57.43%**	83.40%	58.51%	90.89%
4	91.86%	50.48%	83.56%	56.64%	90.72%
5	92.30%	56.29%	83.25%	58.06%	90.42%
6	91.74%	56.58%	82.20%	55.30%	89.46%
7	91.87%	56.68%	83.45%	58.56%	**90.90%**
11	**92.90%**	55.87%	**83.80%**	56.20%	90.60%
mean	92.23%	55.56%	83.27%	57.21%	
Reference EMD^2 models					
11	91.87%	42.06%	81.56%	46.64%	89.43%
11 w	89.31%	51.84%	72.34%	59.86%	82.99%

EMD² IceDisc

Similarly to the CE IceDisc combinations, we see the CE IntDisc scoring lowest on the mean scores and the MSE IntDisc multistage inferences scoring the best closely followed by the BCE multistages. It is also noticeable that the highest individual scores are wider distributed among the IceDisc/IntDisc combinations.

2-class IceDisc

Comparing the 2-class IceDisc/IntDisc multistages to the equivalent means of the multistages with additional classes, it is apparent that they score worse and with a MacroBins score lower than the reference weighted models, which is also lower than the other models.

CE vs EMD² IceDiscs

Generally, it seems that ice discrimination with EMD2 performs better than CE equivalent models in terms of both the ordinary R^2 and the intermediate version. However, in terms of both the ~~multi-accuracy~~ flexible-class accuracy and the intermediate version, the CE multistage inferences take slight precedence in terms of the average scores. Furthermore, in contrast to the CE IceDisc model combinations, the equivalent EMD^2 models have the highest performances somewhat more spread out in between IceDisc/IntDisc model combinations, whereas the CE IceDisc versions are exactly the same across the board.

Comparison to reference models

Inspecting the 11-class reference CE and EMD^2 models, we see substantially lower intermediate R^2-scores, and both ~~multi-accuracy~~ flexible-class accuracy scores compared to the multistage inferences, except for the 11-class CE IceDisc combinations. Both IceDisc combinations with the regression-based IntDisc models have scores that ~~exceed~~ exceeds the reference weighted optimised models in terms of the intermediate R^2-score, though the intermediate ~~multi-class~~ flexible-class accuracy is best for the weighted reference model, though several multistages approach this score.

General remarks

Overall, we see the largest differences in performances across different IceDisc and IntDisc model combinations rather than the changing number of classes the IceDisc model is trained on. Despite varying numbers, it is clear that the IceDisc model shapes what the combined model excels at, e.g. top performances are almost found for the same ice disc models across the int disc models.

4.4 Qualitative multistage inference analysis

For the qualitative assessment, 4 scenes are selected from the test scene subset and are illustrated in Fig. 2-5. Each subfigure includes the HH and HV SAR images, the professionally labelled SIC chart (ground truth), and the associated multistage and reference model outputs highlighted in Tab. 5 with red boxes. For simplicity, we show only three different multistage inferences for each IceDisc loss function but with alternating and no overlap between the number of classes. In addition, multistage inferences are selected based on high performance. The number of classes for the IceDisc models is repeated for convenience in addition to the R^2-scores for the scene. Scenes in Fig. 3-5 were selected to showcase examples where models had relatively diverging outputs measured by the standard variation in R^2-scores across all multistage inferences.

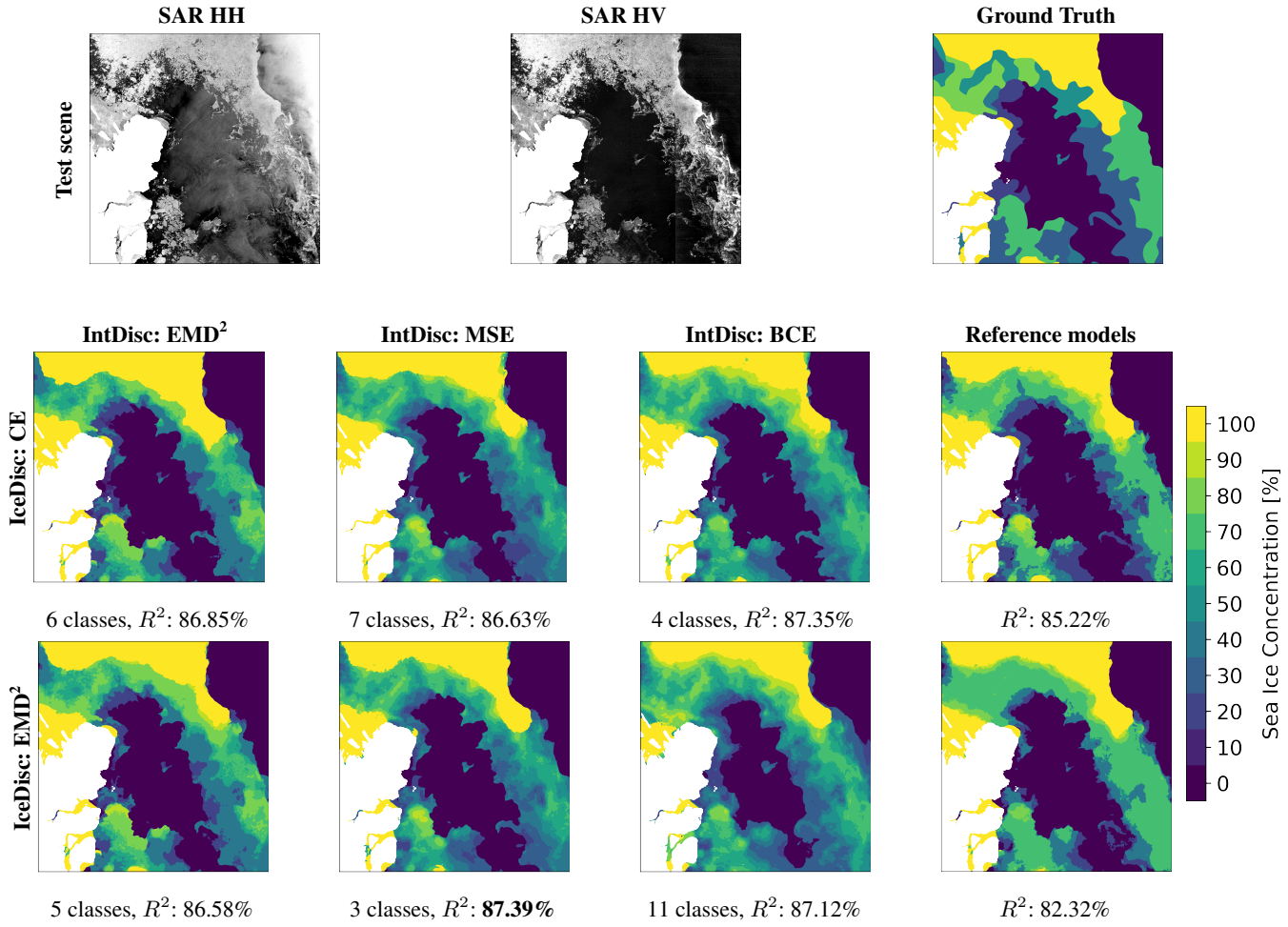


Figure 2. Fram Strait, Northeast Greenland. Scene acquired on September 3, 2018. The top row shows the test scene with the SAR HH and HV images and the corresponding ice charts. Second row: IceDisc Ensemble and reference models using the ~~CrossEntropy~~ CE loss function, and the third row with the ~~Earth Mover's Distance squared~~ EMD^2 loss function. Columns 1-3 have different loss function optimised IntDisc models. Column 4 shows the reference model outputs. All model outputs are shown with the number of classes used to train the IceDisc model and the associated R^2-score.

The scene in Fig. 2 was previously presented in Fig. 1. All 8 models produce outputs resembling the ground truth with clear 0% and 100% ice boundaries. While the 3-class EMD^2 with the MSE IntDisc attains the highest R^2-scene-score, all the multistage combinations score similarly with less than 1% performance difference from best to worst. In addition, the multistage inferences also all score better than the reference models by 1-5%. Furthermore, there is a clear discrepancy between the classification and regression-based IntDisc models with frequent transitions among intermediate classes not visible in the human-labelled ice chart. In addition, the EMD^2 IceDisc and BCE IntDisc combination appear with less sharp open water and ice boundaries compared to the remaining models.

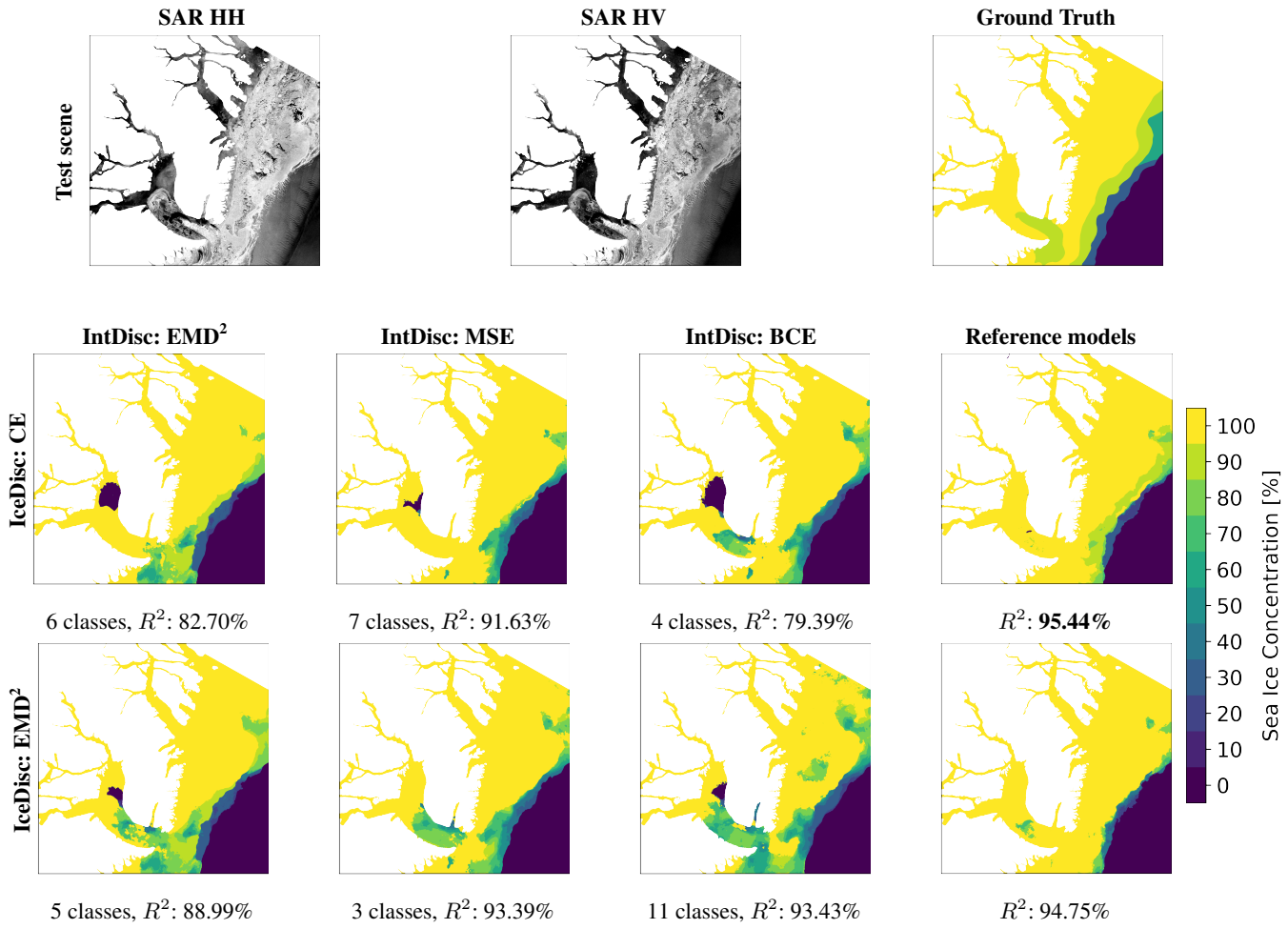


Figure 3. Scoresbysund, East Greenland. Scene acquired on March 15, 2018.

The scene in Fig. 3 was acquired in March 2018 from the Scoresbysund fjord in Eastern Greenland. Sea ice is present along the coast with a high concentration close to the coast and lower concentrations towards the ice edge. There are multiple areas with fast ice, both in the main fjord — Scoresbysund — but also to the North of it, as indicated by the dark SAR signatures with some ice breakup textures. All the models capture the large accumulation of ice along the coast accurately as well as the lower concentrations near the edge. The best-performing multistage inferences are the EMD^2 IceDisc with the MSE or BCE IntDisc combination (0.04% discrepancy). However, both reference models outperform the multistage inferences due to their superior performance (up to 16% better) in the Scoresbysund fjord with fast ice. The additional fast ice in the small fjords is, however, predicted correctly by all models.

In Fig. 4 we present a summer scene during June 2018 from the South of Scoresbysund in the Greenland Strait between Greenland and Iceland. Here, turbulent waters mix ice and water forming swirling textures. The scene generally has low SICs but interesting SAR signatures with a near-range field in the left portion of the image and with some wind in the mid and lower

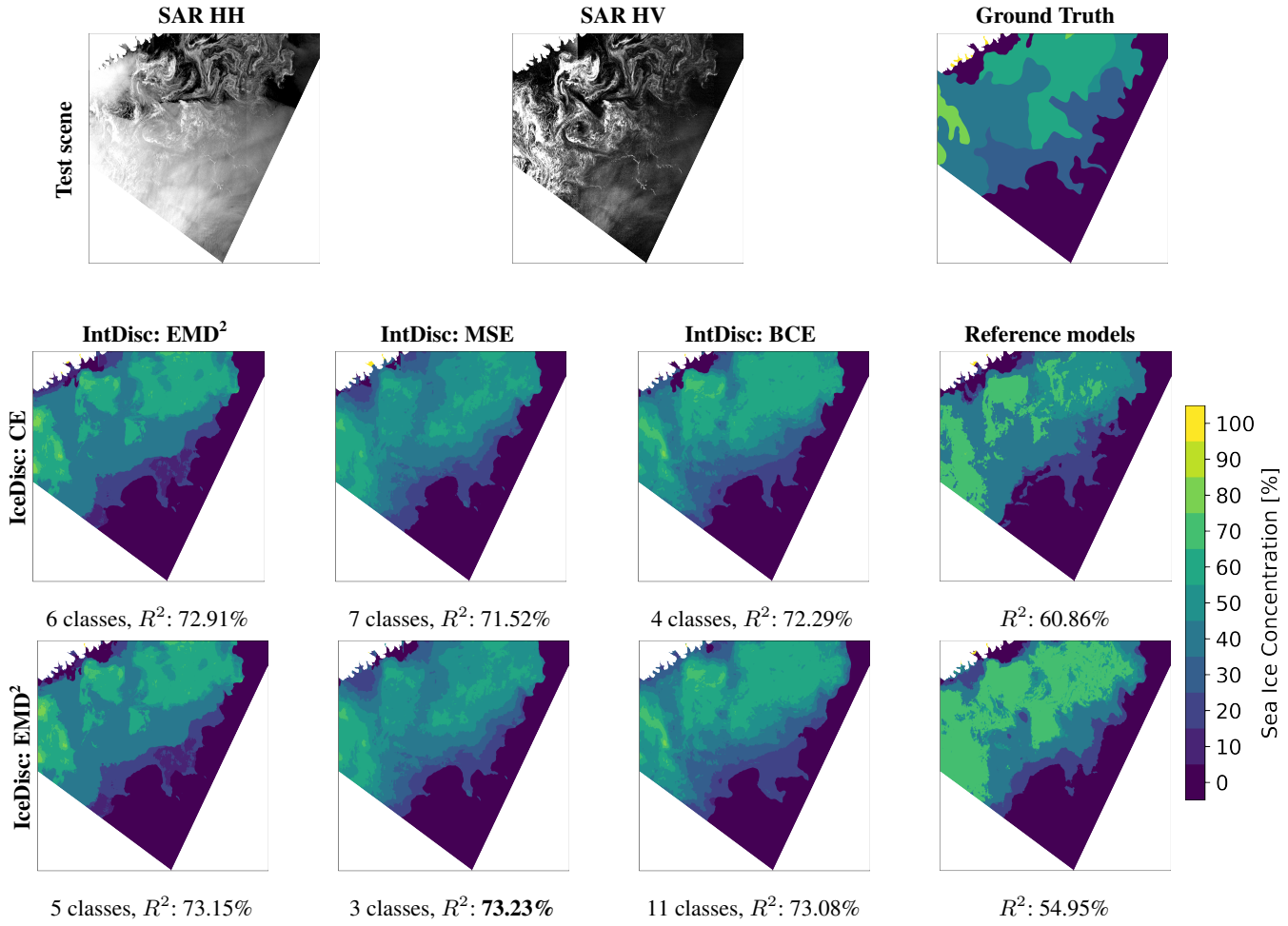


Figure 4. Greenland Strait, between Greenland and Iceland. Scene acquired on June 7, 2018.

part. Overall, the models produce outputs that resemble the human-labelled chart well with sharp edges between the ice areas and open water. The best performance is again achieved by the EMD^2 IceDisc and MSE IntDisc combination, although the performances between the multistage inferences are very close (up to 1.7% discrepancy). On the other hand, both reference models score significantly lower (11-18%) due to the assignment of higher SIC compared to the human-labelled ice chart in the interior of the ice pack. It is also more noticeable that there is a large overlap between the multistage inferences utilising the same IntDisc as in this scene the majority of the pixels contain intermediate SICs.

Fig. 5 illustrates another summer scene from June 2018 but from Baffin Bay in Northwest Greenland, South of Qaanaaq. The scene contains a large area with an abundance of mostly small broken floes pushed together with a couple of larger floes scattered about. The scene consists of a number of low to intermediate ice concentrations and a large area of 90% ice. The models illustrate good overlap with the human-labelled ice chart in terms of the intermediate SICs and open-water areas. Again, the best performance is obtained by the EMD^2 IceDisc and MSE IntDisc combination. In addition, the multistage inferences

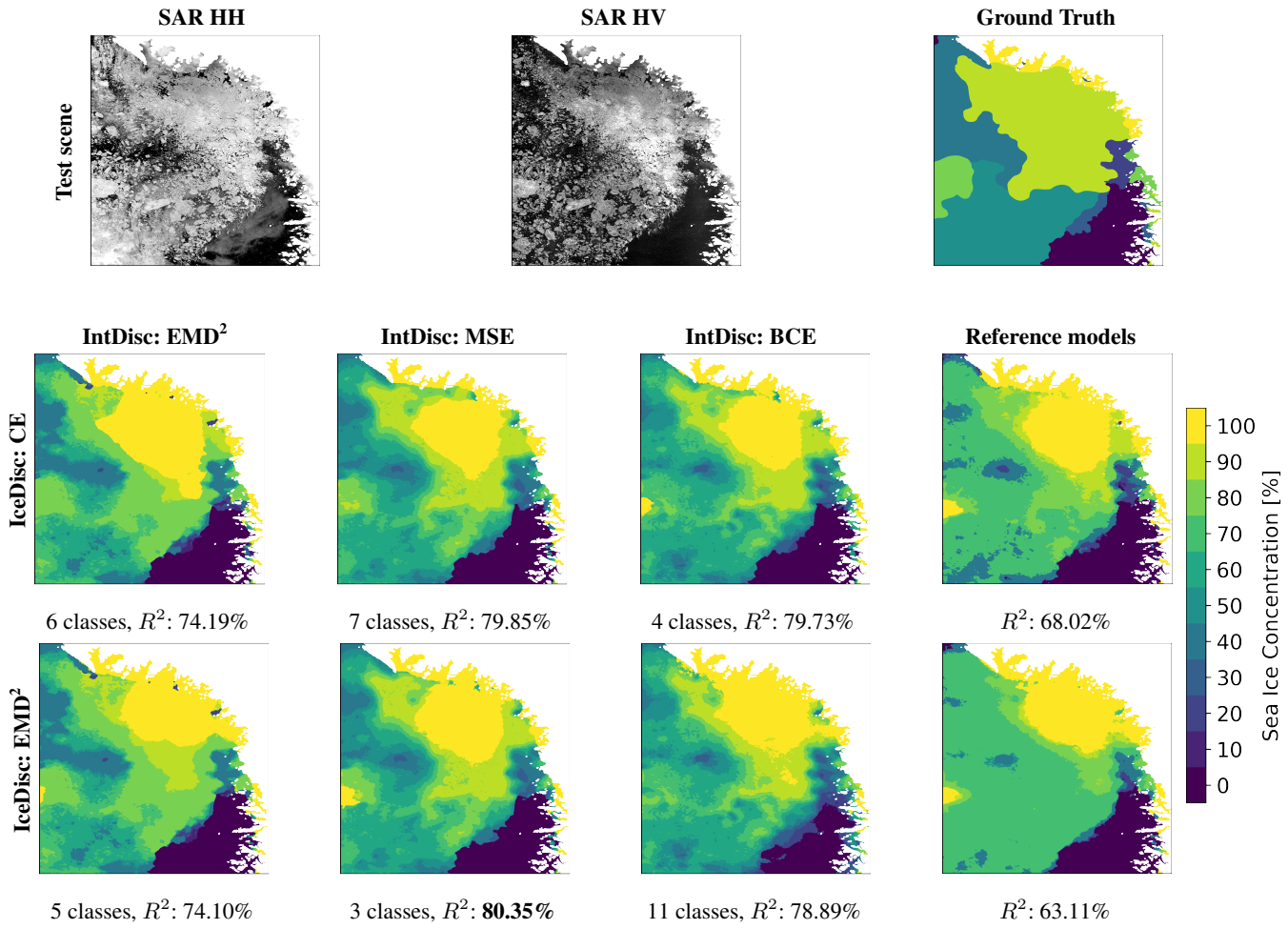


Figure 5. Baffin Bay, Northwest Greenland. Scene acquired on June 27, 2018.

again do better than reference models (6-17%), which provide outputs that are much less nuanced. Interestingly, all the trained models are in disagreement with the human-labelled ice chart with respect to the concentration of the most densely packed ice area. The models label it as 100% whereas the ice chart was assigned 90% but covers a much larger area. It could be argued that this is a case where the SIC depends on how wide or large the polygon is drawn as there is definitely an area that does not contain much water in between the smaller floes.

5 Discussion

From Tab. 3 we see that IceDisc models with fewer classes (3-7) are slightly better with a maximum improvement of 0.73% for CE Icediscs and 0.3% for EMD^2 IceDisc in separating the macro class categories compared to regularly trained 11 class models selected based on the MacroBins metric. However, the superior separability usually stems from an increase in scoring for the

intermediate pixels and is associated with a decline in open water and 100% SIC performance. It is somewhat surprising that separating pixels into fewer categories does not lead to greater separability improvements, as logically, it should be an easier task to predict fewer classes. EMD^2 has some degree of inter-class understanding as it calculates the difference between the assigned individual class probability distribution and the reference distribution. However, It is evident that this additional inter-class relationship capability of the EMD^2 over the CE loss function does not have a significant impact on the model's ability to separate the macro classes.

For the IntDisc models we see a clear difference in the R^2-score listed in Tab. 4 between the classification and regression optimised models with a clear preference for the latter. Given the strong inter-class relationships of the intermediate classes, it is natural that regression-based loss functions excel at reconstructing these classes, as MSE and BCE model the loss based on a geometrical distance and a general ice probability, respectively. The CE and EMD^2 loss functions on the other hand have a lower capability of modelling the inter-class relationships with the CE assigning individual class probabilities. Here, however, the additional inter-class capability of the EMD^2 loss does appear to have a superior influence over the CE loss. With the degree of inter-class relationships ranging from none to most, CE, EMD^2, BCE and MSE, are in the same order of the IntDisc R^2 performance. In effect, these numbers present the best-case performance for the intermediate pixels, as once it is combined with the IceDisc model some true intermediate pixels will be misclassified as either open water or 100% SIC since the *int in int* score ranges from 79-86% leaving 14-21% less intermediate pixels to be correctly classified. Therefore, improving macro class separation will directly benefit the overall intermediate class predictions, and it could therefore be argued that further developments in this direction should take precedence.

The multistage inferences generally score better than the reference models in Tab. 5 except in relation to 0% and 100% SIC as noted earlier and in the case of the mean R^2, where the CE/EMD^2 IceDisc and CE/EMD^2 IntDisc combinations score poorer or on par with the reference models. Comparing the CE and EMD^2 IceDisc combinations, the multistage inferences' achievements on the intermediate R^2-score see an average improvement of 6.87%-20.42% and 3.79%-16.43%, respectively, for the multi accuracy the range is 0.19%-1.41% and 0.57%-1.81%, respectively, and similarly, the intermediate multi accuracy performance difference is 1.97%-6.02% and 6.12-10.9%, respectively. Thus in these terms, there are clear advantages to utilising the multistage inference combinations from an overall numerical perspective. In fact, some of the inference combinations are achieving intermediate SIC performances with respect to the intermediate R^2-score and the intermediate multi-class accuracy that resembles and in some cases exceeds the class-weighted optimised model scores. The class-weighted optimised models are trained with a large emphasis on the intermediate classes, and allowing for a similarly strong performance on the intermediate classes while retaining high MacroBins scores is promising.

With respect to the 2-class IceDisc and Int100Disc inference combination, we see that this approach does not lead to improvements compared to the other trained models, and thus we cannot advise utilising this strategy as presented here. This is due to the low separability measured in MacroBins stemming from the low performance in classifying 100% SIC correctly. Classifying open water and ice can be problematic as the ice polygons utilised as the reference contain both water and ice, and thus the degree of separation between the binary classes is deficient for a problem defined in this setting.

Qualitatively inspecting Figs. 2-5, the multistage inferences perform well and outperform the reference models in 3 out of 4 scenes. The scene where the reference models take precedence is Fig. 3, in which a large area in the Scoresbysund fjord is covered by fast ice. The fast ice is misclassified by the IceDisc models but correctly classified in the reference models. These areas are inherently difficult to classify as they represent an evident ambiguity between open water and sea ice. In practice, sea ice analysts will usually identify these areas of fast ice by using their experience and knowledge of the area, looking at time series of image acquisitions or validate their assumptions through other sensors or communicating with people who can verify the situation in situ. Therefore, it can arguably be optimistic to expect models trained solely on SAR data to reliably correctly identify such fast ice areas. Nevertheless, 100% SIC pixels in areas of fast ice may explain a significant portion of the 100% accuracy discrepancy between the multistage inferences and reference models.

Another discrepancy between the multistage inferences and both the reference outputs and the human-labelled ice charts is the degree of the transition between classes in the best-performing models, which are using regression-optimised IntDisc models. These high frequencies of transition are not present in the original ice chart, so despite greater numerical similarity, visually, it can be less comparable. Though some may argue that this is an improvement in resolution as the human-labelled ice charts are much coarser with large polygons. However, it is also a matter of usage of the large detail level from a navigational perspective, as both the ice will move in between SAR image acquisition and the delivery time of the product. Therefore, the high level of detail may be less relevant as ice drift will have occurred. The second aspect is the linked to the ability of providing a quick overview of the different areas in the sea ice, which may add unnecessary complexity. On the other hand, it could be argued that the output maps from the reference models contain too little information e.g. Fig. 2 and 5.

From inspecting Fig. 5 it could be beneficial to modify the multi-accuracy metric to extend the acceptable range between 90% and 100%. This scene is a ~~classical~~ classic example of an area that is fully covered by sea ice but since the drawn polygon is much wider and ~~covering~~ covers areas with some open water, the result is a lower SIC in the ~~human labelled~~ human-labelled ice chart. Experimenting with extending the range of which predictions are deemed correct, from 10% to 20%, as this is more in line with the average uncertainty of 20%.

In Kucik and Stokholm (2023) the highest scoring R^2 models were optimised using the MSE loss function with a reported score of 93.12% and 92.64% for the BCE equivalent. The main downside of these models was the inadequate 0% and 100% ~~performances~~ accuracies with scores of 83.6% and 80.41%, respectively~~. For~~ , for the MSE and 90.3% and 83.76% for the BCE optimised model. These low scores often occurred due to incorrect ~~labelling~~ segmentation of areas with high SAR noise, bright near-range fields or wind-roughened open water areas. Multiple variants of the EMD^2 IceDisc and MSE or BCE IntDisc combinations score similar on the R^2-score but have a significantly improved 0% and 100% scores making this method a promising approach to both having a high separability and good intermediate SIC performance.

However, despite the objective numerical metric improvements it is somewhat disheartening that the multistage inferences are not able to break the 93% R^2-score ceiling in the current setting. As described in 2.2, there is inherent uncertainty in the SIC ice charts — in particular for the intermediate classes. Thus it may not be realistic to expect those models could generate numerical results close to 100%. However, it is somewhat evident that there are still misclassifications that could be resolved such as the mislabelling of dark microwave SAR signatures from fast ice that resemble still or calm open water.

Adding additional data types, such as passive microwave radiometers (PMR) should, in theory, be able to better deal with these obstacles as the passive microwave radiation are less dependent on the surface roughness properties compared to SAR. Arguably the largest obstacle in using PMR data is the much coarser resolution compared with the SAR data, which may cause spillover effects where emissions from land areas enter pixels covering parts of the ocean near the shore. This is most problematic for smaller fjords. Though, it could be possible to circumnavigate this obstacle by either including information regarding the distance to land or simply removing PMR pixels that are too close to the shoreline. ~~Another avenue could involve investigating super resolution~~

6 Conclusions

This study presents how different CNN models specialised in solving partial tasks in assigning Sea Ice Concentration (SIC) using Sentinel-1 SAR images can be combined in model inferences while benefiting from advantages achieved through model optimisation using either classification or regression-based loss functions. To measure model performance, two new metrics are defined; the MacroBins-score measures the average separation percentage of open-water, intermediate SIC, and 100% SIC; and the flexible-class accuracy that measure accuracy while better incorporating the uncertainty associated with the SICs. Two types of models are defined; the IceDisc models for the macro class category separation, and the IntDisc models to assign specific intermediate classes. We investigate whether combining classes will improve the separation of the macro classes. IceDisc models are trained with 7 different class combinations with either the CrossEntropy (CE), or Earth Mover's Distance squared (EMD^2) loss functions. IntDisc models are trained with either classification-oriented loss functions: CE, EMD^2 or regression-based: Mean Square Error (MSE), Binary CrossEntropy (BCE).

Overall, we see the IceDisc/IntDisc multistage inferences being capable of producing SIC maps that resemble those produced by human ice analysts while scoring well on the selected metrics. We note that the best-performing model combinations are achieved by the EMD^2 IceDisc and either MSE or BCE IntDisc inference combinations reaching an R^2-score of $\sim 93\%$ similar to the best scoring models in a previous study by Kucik and Stokholm (2023). However, the multistage inferences achieve significantly improved accuracy for open water areas and areas with 100% sea ice concentration. Compared to the reference models optimised with CE or EMD^2 selected to achieve high R^2-scores, the multistage inferences outperform both total R^2 and multi-class accuracy scores and the intermediate class equivalents. Also, the multistage inferences are inferior in terms of the open water and 100% SIC accuracies, where the latter is negatively impacted in some cases by shortcomings in correctly distinguishing areas of fully covered fast ice exhibiting dark microwave signatures resembling open water. However, these obstacles could be solved by including other data sources such as passive microwave radiometry that is less dependent on the ice surface roughness. Another outcome of this study is the appearance of reducing different numbers of classes from 3 to 7 does not seem to impact the separability of the macro SIC classes. Instead, the largest performance change came from the choice of loss function used in the IntDisc model optimisation.

Thus the conclusion of the study is that combining multiple models specialised in the different tasks presented can lead to the automatic production of SIC maps that have improved intermediate class performance while retaining good accuracy in open water and 100% SIC areas compared to the ordinary reference models optimised with the CE or EMD^2 loss functions.

7 Future Work

Further improvements in the assignment of SIC could be achieved by adding more types of data, such as passive microwave radiometer data, which could better help identify areas of fast ice. Additional data types such as environmental parameters, geographical information or seasonality could also be beneficial. Another option could be to expand the AI4Arctic Sea Ice Dataset with more data on particularly the intermediate SIC classes could be beneficial, which are currently the least abundant classes in the current version (Stokholm et al., 2022). Increasing the distribution of seasonality and regionality could also be beneficial. In this connection, expanding the test dataset would also provide more scenarios in which we can validate the model performance yielding greater confidence in the outputs. Finally, while the U-Net architecture is decent and simple, many new types of CNN and computer vision architectures have been developed, such as ConvNeXt (Woo et al., 2023). These may be interesting to investigate.

Code availability. CNN model architecture and loss function code is available at: https://github.com/astokholm/AI4SeaIce.git (Stokholm and Kucik)

Data availability. The original dataset is available at https://data.dtu.dk/articles/dataset/AI4Arctic_ASIP_Sea_Ice_Dataset_-_version_2/13011134/3 (Saldo et al., 2021)

Video supplement. TEXT

Author contributions. The methodology, experiments, analysis and manuscript writing were implemented and carried out by Andreas Stokholm. Idea formulation, conceptualisation and software development were developed in collaboration between Andrzej Kucik and Andreas Stokholm. The study was carried out under the supervision and support of Nicolas Longépé and Sine Munk Hvidegaard. All authors have reviewed the manuscript.

Competing interests. Authors declare that none of the authors has any competing interests.

Acknowledgements. The authors would like to thank DTU Space and ESA ϕ-lab for their support of this research, and the AI4Arctic team who developed the dataset used. In addition, thanks to the Niels Bohr Foundation and the Thomas B. Thriges Foundation who provided financial assistance during Andreas Stokholm's external research stay at ESA ϕ-lab, where much of this study was formulated and carried out.

References

Bekkers, E., Francois, J. F., and RojasRomagosa, H.: Melting ice Caps and the Economic Impact of Opening the Northern Sea Route, The Economic Journal, 128, 1095–1127, [Online; accessed 2022-10-18], 2017.

Boulze, H., Korosov, A., and Brajard, J.: Classification of Sea Ice Types in Sentinel-1 SAR Data Using Convolutional Neural Networks, Remote Sensing, 12, 2165, [Online; accessed 2023-04-13], 2020.

Boutin, G., Williams, T., Rampal, P., Olason, E., and Lique, C.: Impact of wave-induced sea ice fragmentation on sea ice dynamics in the MIZ, Tech. rep., Copernicus GmbH, [Online; accessed 2023-04-13], 2020.

Cheng, A., Casati, B., Tivy, A., Zagon, T., Lemieux, J.-F., and Tremblay, L. B.: Accuracy and inter-analyst agreement of visually estimated sea ice concentrations in Canadian Ice Service ice charts using single-polarization RADARSAT-2, The Cryosphere, 14, 1289–1310, [Online; accessed 2023-01-23], 2020.

de Gelis, I., Colin, A., and Longepe, N.: Prediction of categorized sea ice concentration from sentinel-1 SAR images based on a fully convolutional network, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14, 5831–5841, [Online; accessed 2022-10-18], 2021.

Heidler, K., Mou, L., and Zhu, X. X.: Seeing the Bigger Picture: Enabling Large Context Windows in Neural Networks by Combining Multiple Zoom Levels, in: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, IEEE, [Online; accessed 2023-04-13], 2021.

Huang, B., Collins, L. M., Bradbury, K., and Malof, J. M.: Deep convolutional segmentation of remote sensing imagery: A simple and efficient alternative to stitching output labels, in: IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, IEEE, [Online; accessed 2022-10-19], 2018.

Karvonen, J., Vainio, J., Marnela, M., Eriksson, P., and Niskanen, T.: A comparison between high-resolution eo-based and ice analyst-assigned sea ice concentrations, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 8, 1799–1807, [Online; accessed 2022-10-18], 2015.

Kasahara, M., Imaoka, K., Kachi, M., Fujii, H., Naoki, K., Maeda, T., Ito, N., Nakagawa, K., and Oki, T.: Status of AMSR2 on GCOM-W1, in: SPIE Proceedings, SPIE, [Online; accessed 2023-01-23], 2012.

Khaleghian, S., Ullah, H., Kræmer, T., Hughes, N., Eltoft, T., and Marinoni, A.: Sea ice classification of SAR imagery based on convolution neural networks, Remote Sensing, 13, 1734, [Online; accessed 2022-10-18], 2021.

Korosov, A., Demchev, D., Miranda, N., Franceschi, N., and Park, J.-W.: Thermal Denoising of Cross-Polarized Sentinel-1 Data in Interferometric and Extra Wide Swath Modes, IEEE Transactions on Geoscience and Remote Sensing, 60, 1–11, [Online; accessed 2023-04-04], 2022.

Koubarakis, M., Stamoulis, G., Bilidas, D., Ioannidis, T., Mandilaras, G., Pantazi, D.-A., Papadakis, G., Vlassov, V., Payberah2, A., Wang, T., Sheikholeslami, S., Hagos, D. H., Bruzzone, L., Paris, C., Weikmann, G., Marinelli, D., Eltoft, T., Marinoni, A., Kræmer, T., Khaleghian, S., Ullah, H., Troumpoukis, A., Kostopoulou, N. P., Konstantopoulos, S., Karkaletsis, V., Dowling, J., Kakantousis, T., Datcu, M., Yao, W., Dumitru, C. O., Appel, F., Migdall, S., Muerth, M., Bach, H., Hughes, N., Everett, A., Kiærbech, A., Pedersen, J. L., Arthurs, D., Fleming, A., and Cziferszky, A.: Artificial Intelligence and Big Data Technologies for Copernicus Data: The ExtremeEarth Project, in: Publications Office of the EU, pp. 9–12, 2021.

Kucik, A. and Stokholm, A.: AI4SeaIce: Comparing Loss Representations for SAR Sea Ice Concentration Charting, https://ai4earthscience.github.io/iclr-2022-workshop/accepted, [Online; accessed 2023-05-10], 2022.

Kucik, A. and Stokholm, A.: AI4SeaIce: selecting loss functions for automated SAR sea ice concentration charting, Scientific Reports, 13, [Online; accessed 2023-04-13], 2023.

Malmgren-Hansen, D., Pedersen, L. T., Nielsen, A. A., Skriver, H., Saldo, R., Kreiner, M. B., and Buus-Hinkler, J.: ASIP Sea Ice Dataset - Version 1, https://data.dtu.dk/articles/ASIP_Sea_Ice_Dataset_-_version_1/11920416/1, 2020.

Malmgren-Hansen, D., Pedersen, L. T., Nielsen, A. A., Kreiner, M. B., Saldo, R., Skriver, H., Lavelle, J., Buus-Hinkler, J., and Krane, K. H.: A convolutional neural network architecture for sentinel-1 and AMSR2 data fusion, IEEE Transactions on Geoscience and Remote Sensing, 59, 1890–1902, [Online; accessed 2022-10-18], 2021.

Park, J.-W., Korosov, A. A., Babiker, M., Sandven, S., and Won, J.-S.: Efficient Thermal Noise Removal for Sentinel-1 TOPSAR Cross-Polarization Channel, IEEE Transactions on Geoscience and Remote Sensing, 56, 1555–1565, [Online; accessed 2023-04-07], 2018.

Park, J.-W., Won, J.-S., Korosov, A. A., Babiker, M., and Miranda, N.: Textural Noise Correction for Sentinel-1 TOPSAR Cross-Polarization Channel Images, IEEE Transactions on Geoscience and Remote Sensing, 57, 4040–4049, [Online; accessed 2023-04-07], 2019.

Perovich, D., Meier, W., Tschudi, M., Hendricks, S., Petty, A. A., Divine, D., Farrell, S., Gerland, S., Haas, C., Kaleschke, L., Pavlova, O., Ricker, R., Tian-Kunze, X., Webster, M., and Wood, K.: Arctic Report Card 2020: Sea Ice, https://repository.library.noaa.gov/view/noaa/27904, 2020.

Radhakrishnan, K., Scott, K. A., and Clausi, D. A.: Sea Ice Concentration Estimation: Using Passive Microwave and SAR Data With a U-Net and Curriculum Learning, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14, 5339–5351, [Online; accessed 2023-04-13], 2021.

Ronneberger, O., Fischer, P., and Brox, T.: U-Net: Convolutional networks for biomedical image segmentation, pp. 234–241, Springer International Publishing, Cham, [Online; accessed 2022-10-18], 2015.

Saldo, R., Kreiner, M. B., Buus-Hinkler, J., Pedersen, L. T., Malmgren-Hansen, D., Nielsen, A. A., and Skriver, H.: AI4Arctic / ASIP Sea Ice Dataset - Version 2, https://data.dtu.dk/articles/dataset/AI4Arctic_ASIP_Sea_Ice_Dataset_-_version_2/13011134/3, 2021.

Stokholm, A. and Kucik, A.: AI4SeaIce Github, https://github.com/astokholm/AI4SeaIce.git.

Stokholm, A., Wulf, T., Kucik, A., Saldo, R., Buus-Hinkler, J., and Hvidegaard, S. M.: AI4SeaIce: Toward solving ambiguous SAR textures in convolutional neural networks for automatic sea ice concentration charting, IEEE Transactions on Geoscience and Remote Sensing, 60, 1–13, [Online; accessed 2022-10-18], 2022.

Tamber, M. S., Scott, K. A., and Pedersen, L. T.: Accounting for label errors when training a convolutional neural network to estimate sea ice concentration using operational ice charts, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15, 1502–1513, [Online; accessed 2022-10-18], 2022.

Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., Potin, P., Rommen, B., Floury, N., Brown, M., Traver, I. N., Deghaye, P., Duesmann, B., Rosich, B., Miranda, N., Bruno, C., L'Abbate, M., Croci, R., Pietropaolo, A., Huchler, M., and Rostan, F.: GMES Sentinel-1 mission, Remote Sensing of Environment, 120, 9–24, [Online; accessed 2023-01-23], 2012.

Wang, L., Scott, K. A., Xu, L., and Clausi, D. A.: Sea ice concentration estimation during melt from dual-pol SAR scenes using deep convolutional neural networks: A case study, IEEE Transactions on Geoscience and Remote Sensing, 54, 4524–4533, [Online; accessed 2022-10-18], 2016.

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, [Online; accessed 2023-04-13], 2017a.

Wang, L., Scott, K. A., Clausi, D. A., and Xu, Y.: Ice concentration estimation in the gulf of St. Lawrence using fully convolutional neural network, in: 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), IEEE, [Online; accessed 2023-04-13], 2017b.

Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I. S., and Xie, S.: ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders, https://arxiv.org/abs/2301.00808, 2023.