Model Ensemble with Dropout for Uncertainty Estimation in Binary Sea Ice or Water Segmentation using Sentinel-1 SAR

Rafael Pires de Lima, Morteza Karimzadeh

***Abstract*—Despite the growing use of deep learning in sea ice mapping with SAR imagery, the study of model uncertainty and segmentation results remains limited. Deep learning models often produce overconfident predictions, a concern in sea ice mapping where misclassification can impact marine navigation safety. We incorporate and compare dropout and model ensemble within a convolutional neural network segmentation architecture to highlight regions with prediction uncertainty, and explore the impact of loss function choice. We evaluate model generalization and uncertainty characterization by training and evaluating models on the AI4Arctic Sea Ice Challenge Dataset (primary). We further explore model uncertainty by testing the trained models on the Extreme Earth version 2 Dataset (secondary). The primary and secondary datasets vary in number of scenes as well as in the available data and preprocessing. We obtain test F1 results higher than 0.97 for the primary dataset. Although the F1 performance for the secondary dataset is reduced to 0.93, the generated sea ice maps are reasonable across several Sentinel-1 scenes, and our proposed strategy helps in identification of misclassified and uncertain regions for human quality control. Our models seem to be robust against banding noise in Sentinel-1 SAR, and the prediction uncertainty frequently highlights ice regions misclassified as water, indicating its potential for real-world applications. Our study advances the field of machine learning-based sea ice mapping and highlights the importance of uncertainty estimation and cross-dataset evaluation for model development and deployment. Our approach can be adopted for other remote sensing applications as well.**

***Index Terms*—Arctic, convolutional neural networks, image segmentation, SAR, sea ice.**

I. Introduction

DEEP LEARNING, which has revolutionized many fields of research and practice, is increasingly being utilized in remote sensing. However, its performance varies based on the sensor used and the object being studied. One particularly challenging area for deploying deep learning is sea ice mapping using Synthetic Aparture Radar (SAR) imagery.

The increase in average global temperatures and declining sea ice extent has led to a rise in the number of ships that navigate the Arctic [1]–[3]. Ford et al. [4] remarked that the combination of more ice-free open water summers, and the substantial reduction in sea ice volume, in both its extent and thickness, is extending the shipping season in the Arctic Ocean. The decline in sea ice might not only promote an increase in maritime goods transportation, but to make cruise tourism an opportunity for economic development (e.g., [5]). This activity boost requires an increased awareness of ocean conditions and environmental monitoring to guarantee safety and security of marine vessels, and to work towards minimizing their environment impact. Thus, national ice centers of countries with interests in the Arctic that produce crucial information for navigating ice-infested waters [6] will face an increased demand.

National ice centers periodically produce ice charts (i.e., maps) that provide information about the occurrence of ice in both Arctic and Antarctic waters. The degree of detail, as well as the frequency with which these ice charts are generated, vary according to specific objectives. However, a factor that remains present in the production of most ice charts is that they rely on the human expert sea ice analysts' interpretation of remotely sensed data, including SAR as a primary data source. Sea ice charting is labor-intensive. Ice analysts need to manually delineate the extent and identify ice characteristics for large regions in a limited timeframe, as sea ice is highly dynamic due to the combined effect of wind, ocean currents and temperature. SAR, while providing a high-resolution data source that is independent of light and cloud coverage, introduces additional complexities due to ambiguous and overlapping backscatter values for different conditions of sea ice in SAR, increasing uncertainty and time required for interpretation. Different types of sea ice (and at times water) appear visually similar in SAR imagery, a phenomenon caused by several factors, including the salinity of sea ice and surface air pockets. These imaging variations complicate object identification in SAR.

Although the development of automated methods for sea ice characterization can be traced back at least to the late 1980s when Holt et al. [7] used unsupervised algorithms to segment airborne SAR imagery, the recent adoption of Convolutional Neural Networks (CNNs) has accelerated the use of machine learning in the field. The structure of CNNs makes them particularly successful in processing structured data that exhibit organized patterns, such as those imaged by remotely sensed imagery (e.g., [8]–[10]). Even at earlier studies (e.g., [7]), researchers were aware that regional and seasonal variations in sea ice conditions posed a challenge to automated algorithms. Understanding how larger datasets can have a positive effect in the development of machine learning

This work was supported by the National Science Foundation under Grant No. 2026962. (Corresponding author: Rafael Pires de Lima). The code we developed for this project is available at https://github.com/geohai/sea-ice-binary-ai4seaice.

Rafael Pires de Lima is with the Department of Geography, University of Colorado Boulder (e-mail: rlima@colorado.edu)

Morteza Karimzadeh is with the Department of Geography, University of Colorado Boulder (e-mail: karimzadeh@colorado.edu)

algorithms, the sea ice community is also working to produce tailored labeled datasets (e.g., [11]–[13]).

Although the detailed characterization of ice type and concentration is valuable information for many national ice services, fully automated binary classification of open water and ice still remains challenging. Wind (e.g., [14]), wave dampening (e.g., [15]), dark smooth ice, and other factors are manifested in the SAR backscatter and make the identification of ice a challenging task for automated algorithms. One path to move towards partial automation is to generate a preliminary ice-chart for the human-expert analyst to quality control before publication. For this to be time-efficient, ideally, the algorithmic output could also identify areas of uncertainty, where the expert can focus their attention for approving or modifying the models' predictions.

Additionally, models trained using specific datasets may not generalize well to datasets with different preprocessing, which is the case for sea ice mapping, where the (usually small) training datasets may have different characteristics than images ingested during operational phase. Given the paucity of benchmark sea ice datasets, the generalizability of deep learning-based sea ice mapping has not been fully studied. However, with the recent publication of different benchmark datasets, there is new opportunity for investigating the generalizability of these methods, which might provide better indications of the performance if these models were to be deployed, as well as providing a more realistic testbed for methods to characterize uncertainty.

To address these research gaps, in this paper, we incorporate dropout and model ensemble in a CNN segmentation architecture to characterize uncertain predictions, compare their performance, and analyze the effect of loss function on the results. We explore how the models' overconfidence and prediction uncertainty can either facilitate or impede further quality control by a sea ice analyst. Additionally, we adopt a novel approach for the evaluation of model generalization by training models on a primary benchmark dataset and testing on a secondary benchmark dataset. We do so through the development of a CNN-based architecture that performs image segmentation with the objective of discriminating open water and ice. We perform experiments using only SAR data as input.

The main contributions of this paper are:

- We investigate dropout and model ensemble techniques and their combination in a semantic segmentation CNN framework to characterize uncertainty in deep learning applied to SAR imagery, which is useful for human quality control and increased reliability of machine learning-based products for sea ice mapping.
- We train models using two popular classification losses and evaluate their performance for sea ice mapping, and investigate the resulting uncertainty for each loss function choice.
- We evaluate the generalizability of models and our proposed method for characterizing uncertainty on unseen data by testing on a different benchmark dataset. This test dataset was labeled by sea ice analysts from another national ice center, and has a different pre-processing pipeline when compared to the dataset used for training. To the best of our knowledge, this is the first paper that investigates the performance of a model trained on one benchmark dataset and tested on a secondary benchmark dataset for sea ice characterization.

II. Related Work

We first provide a quick overview of related research in sea ice mapping automation, and then, cover related methods in characterizing uncertainty in deep learning output.

There has been growing interest in developing deep learning models for the characterization of sea ice using remotely-sensed data. Most applications fall into one of two categories: patch classification or regression; or semantic segmentation. For patch classification or regression, a sub-image patch is extracted from the input image and assigned a single label or value for the entire patch. For semantic segmentation, the output labels generally have the same dimension as the input image, although the dimensions of the image throughout segmentation may be manipulated, using upscaling or downscaling techniques.

Wang et al. [16] used 11 RADARSAT-2 images of the Beaufort Sea interpreted by sea ice analysts from the Canadian Ice Service to estimate sea ice concentration at the patch level as a regression task. The images were acquired during the melt season between July and September in 2010 and 2011. Wang et al. [17] used 25 RADARSAT-2 images from the Gulf of Saint Lawrence on the East coast of Canada to train CNNs and fully connected neural networks to estimate sea ice concentration using SAR patches. The images were acquired during the freeze-up season, from January to February 2014. They found CNNs models had better performance than the densely connected neural networks. Instead of using interpreted ice charts, Cooke and Scott [18] used lower-resolution passive microwave images to help train a CNN model that uses SAR data as input and outputs sea ice concentration. Passive microwave data are commonly used to estimate sea ice concentration, however spatial resolution is generally in the order of tens of kilometers. Their work can also be interpreted as a patch regression task as the output of the models corresponds to several pixels from the input. Stokholm et al. [19] used the AI4Arctic/ASIP v2 (ASID-v2, [11]) benchmark dataset to estimate sea ice concentration developing new variations of U-Net [20], a semantic segmentation architecture. Kucik and Stokholm [21] showed that the choice of loss function can significantly affect the appearance of the sea ice concentration predictions generated by CNN models.

Continued effort has been put into developing CNN models for the binary classification of SAR image as ice or water. Khaleghian et al. [22] used 31 Sentinel-1 images acquired north of the Svalbard archipelago in winter months between September and March of 2015 to 2018. They classified patches of different sizes as ice or water and found that the additive system noise in the SAR imagery was challenging for their models. Song et al. [23] used ASID-v2 to perform semantic segmentation of ice or water. They developed a model architecture based on the Pyramid Scene Parsing Network [24]

with adaptions to accommodate the uncertainty of ice-water edge. Instead of using entire scenes for testing, [23] cropped patches of size 800 by 800 pixels and selected 10% of those crops to be part of the test set. This can have implications for evaluation, as parts of the image whose crops are used in testing might have already been seen during training. Their model achieved an accuracy of 0.942 and F1 of 0.930. They found a region of misclassification over open water associated with subswath banding effect when evaluating their model on a full Sentinel-1 scenes. Zhang et al. [25] used a model architecture that is also based on the ASPP module. However, instead of satellite data, they were interested in performing semantic segmentation using data from video cameras of a Chinese ice-strengthened cargo ship. They found their model was able to generate precise ice floe boundaries using the video camera data.

An aspect not thoroughly explored in most research on developing CNN-derived sea ice products is the natural stochastic variation of the models, as well as measures of uncertainty. Boulze et al. [26] showed raw output probabilities from their models as a measure for uncertainty. They also highlighted that the raw output probabilities very likely underestimate the models' uncertainty, as was observed by [27], and that different methodologies could be implemented for a more robust uncertainty estimation. Asadi et al. [28] investigated incorporating uncertainty in the context of sea ice products, considering uncertainty both in the model parameters and input features. They observed that adding uncertainty disturbances in their model reduced both the accuracies and the misclassification rates - Asadi et al. [28] considered a pixel as water or ice only when the probability was equal to or higher than 0.95, pixels with intermediate probabilities were considered "unknowns". Guo et al. [29] remarked that modern neural networks are poorly calibrated, meaning they produce overconfident predictions. Such model overconfidence might be related to Asadi et al.'s slight decrease in accuracy at the cost of improved misclassification rates; the models with uncertainty became more calibrated and generated more predictions with smaller probability. Chen et al [30] recently observed that investigating the prediction uncertainty is helpful to better access the reliability of machine learning predictions for binary sea ice classification. They used a Bayesian CNN followed by iterative region growing with semantics [31], [32] and uncertainty value thresholding to classify patches of sea ice and water. Their algorithm allows for the separation of aleatoric (or data) uncertainty and epistemic (or model) uncertainty. Using 21 RADARSAT-2 scenes, they conclude that the primary source of uncertainty is aleatoric, which arises from the large variability in the patterns of ice and water under varying conditions.

There are other alternatives that help us estimate the uncertainty of machine learning models. For example, Gal and Ghahramani [33] proposed using Dropout as an approximation to Bayesian inference in deep Gaussian processes, a methodology sometimes named Monte Carlo Dropout. Dropout [34] randomly deactivates neurons during training, which helps models avoid overfitting by allowing different neural paths to learn useful weights, rather than the accumulation of large weights in only a subset of parameters. Typically, dropout is disabled during test time and the output of a trained model is deterministic. Gal and Ghahramani's methodology maintains dropout enabled during test time. In doing so, the model generates different outputs from the same input. Lakshminarayanan et al. [35] observed that dropout may be interpreted as an ensemble model combination, as the outputs of a model are averaged over an ensemble of models with shared parameters. Dechesne et al. [36] used Monte Carlo Dropout to estimate the uncertainty of U-Nets trained to segment four different datasets representing various urban scenes.

There are other alternatives for uncertainty estimation. For example, Lakshminarayanan et al. [35] proposed an uncertainty estimation method that relies on the combination of model ensembles and adversarial training [37], [38]. Ensemble is used to estimate model uncertainty by averaging predictions over multiple models, while adversarial training encourages local smoothness. Mehrtash et al. [39] studied ensembles for confidence calibration of CNN models trained to segment medical images. They concluded that ensembles were successful for confidence calibration of CNN for model segmentation trained with Dice loss [40]. Dice loss was developed to use Intersection over Union (IoU) to better deal with class imbalance.

III. Datasets and Targets

A. Primary dataset: AI4Arctic Sea Ice Challenge Dataset

We select the AI4Arctic Sea Ice Challenge Dataset version 2 [13], hereinafter "primary dataset", to train our models. To our knowledge, this is the largest openly accessible labeled sea ice dataset at the time of this writing. The primary dataset is composed of 533 scenes that are available in both "raw" and "ready to train" versions. We use the "ready to train" version that has already been pre-processed for denoising Sentinel-1 images using the algorithm described in [41] and resampled to a 80x80m pixels. The images in the dataset cover the Canadian Arctic and the waters surrounding Greenland. Of the 533 scenes, 20 scenes were separated to be part of the challenge test and their labels were not released at the time of this writing. Therefore, we have 513 scenes with labels available for this study. Each one of the 513 scenes contains ice charts produced by either the Greenland Ice Service or the Canadian Ice Service, and a Sentinel-1 image paired with the ice chart.

In our models, we use the pre-processed Sentinel-1 Horizontal-Horizontal (HH) and Horizontal-Vertical (HV) polarizations, as well as the incidence angle of the images as input. In the scenes of the dataset these images are called nersc_sar_primary, nersc_sar_secondary, and sar_incidenceangle, respectively. HH and HV represent the backscatter coefficient (σ_0) in dB, and the incidence angle is measured in degree. The "ready to train" version scenes contain normalized values. Information about mean and standard deviation values used for normalization can be found in [13]. Although the scenes contain information that can be used to do georeferencing, the scenes themselves do not have geographic projection as distributed. We chose to show all images in this

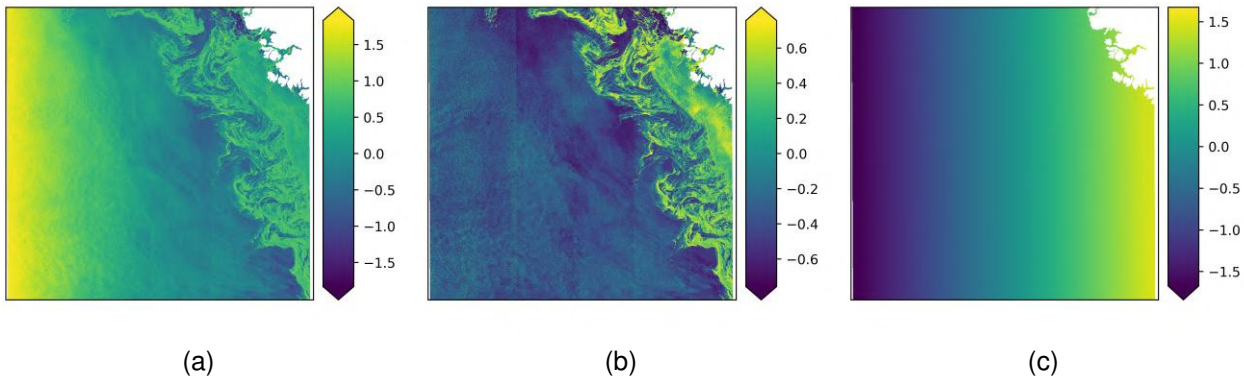


Fig. 1. An example of Sentinel-1 image from scene 20200424T101936_cis_prep.nc in the primary dataset. (a) shows HH, (b) HV, and (c) incidence angle. The top right corner land is masked in white. The colorbars indicate with a semi-arrow which part of the value range was clipped for better display. When clipped, images are in the [2,98] percentile ranges. Each pixel represents 80m^2 and the image is roughly 400 x 400 km.

Fig. 2. Label for the sample 20200424T101936_cis_prep.nc, corresponding to the SAR contents scene in Fig. 1

manuscript using the Sentinel-1 indexing in the dataset, that uses sar_lines to indicate azimuth and sar_samples to indicate range with respect to the satellite. Fig. 1 shows an example of one of the scenes available as input for training. We provide the location of 20200424T101936_cis_prep.nc in Fig. 1, as well as the remaining maps locations in the Supplementary Material.

The "ready to train" version of the primary dataset provides information about sea ice concentration, stage of development, and floe size as labels. To create a binary label that indicates ice or water, we simply selected any regions containing labels for ice concentration equal to or higher than 10% to represent the existence of ice. We label as water regions where the sea ice concentration is zero. The training "ready to train" version of the dataset sums up to 56 GB. Fig. 2 shows the corresponding labels for the Sentinel-1 data in Fig. 1. By converting the labels to ice or water, the 513 scenes in the dataset sum to a collection of labels in which 57% of the pixels are labeled as water and the remaining 43% as sea ice.

Before starting the experimentation, we selected 10 scenes to be held out during training as test samples and to better help us gauge the performance of our models. We selected the test set scenes considering images that had 70% to 80% of pixels labeled for stage of development, with some variation on the ice type. During initial experimentation, we found 12 scenes that had less than 10% pixels labeled, and we chose to discard such scenes. The list of scenes names for both the test selection and the images discarded are provided in the Appendix A. During experimentation, we held out the test set for final evaluation and performed most of the analysis using the validation scenes that will be further discussed in Section V-A. Intensive analysis of the test set in small datasets increases the risk of overfitting the model to the test set, even if the data itself was never used for training. By limiting our analysis on the test set, we can be more confident that the performance of the models on the test set is closer to what would be observed if our model was deployed to process operational data with the same characteristics as the training set.

B. Secondary dataset: ExtremeEarth v2

The ExtremeEarth v2 dataset contains high-resolution sea ice charts specifically labeled for training automated algorithms. It comprises 12 Sentinel-1 images in Extra Wide (EW) mode, acquired over the East Coast of Greenland throughout 2018, intentionally selected by dataset creators to cover diverse types of sea ice and varying weather conditions, and thus, our choice for evaluation of generalizability of the models. Expert ice analysts at the Norwegian Meteorological Institute (MET Norway) drew polygons and assigned sea ice properties primarily based on the interpretation of Sentinel-1 backscatter signatures, along with other remote sensing data. The interpretation was performed with a higher granularity than that typically used for operational sea ice mapping, resulting in smaller polygons than what is typical used for daily or weekly sea ice charts. Sea ice analysts assigned to each polygon values for total sea ice concentration, primary sea ice type (the oldest ice in the polygon, coded SA in the Egg Code), secondary ice type (the second oldest, coded SB), and partial concentrations for each identified ice type. Moreover, they assigned floe size for each ice type, as well as identified whether icebergs were

present. ExtremeEarth v2 is defined hereinafter "secondary dataset" and we use it as the main test dataset.

To prepare the Sentinel-1 images for our models, we preprocessed the data using SNAP version 8. We used the following steps: reprojection, thermal noise removal, orbit file correction, radiometric calibration, speckle filtering, and terrain correction. Hughes and Amdal [12] observed that terrain correction is important for correctly geolocating ascending Sentinel-1 images on the east coast of Greenland. Additionally, we resampled the Sentinel-1 scenes from 40 by 40 meters pixel spacing to 80 by 80 meters pixel spacing. To convert the raw backscattered values of HH and HV bands to decibels, we used GDAL [42], which resulted in a distribution of HH and HV data that is closer to a Gaussian-like distribution, and conforms to the dB transformation in the primary dataset. For testing the secondary dataset, we use the same mean and standard deviation from the primary dataset.

IV. Methods

We first describe the model architecture and training strategy to contextualize our approach for characterizing uncertainty in a replicable manner. We then describe the training and testing split. Finally, we describe the uncertainty methods used in this study.

A. Model Architecture

Fig. 3 shows the model architecture. The model encoder uses the first two residual blocks from ResNet-18 (i.e., up to "layer2"). The decoder module is based on Atrous Spatial Pyramid Pooling (ASPP), which processes feature maps at various scales using atrous convolutions to capture regional spatial context important for sea ice characterization. The original ASPP module employs global average pooling and interpolation for global context incorporation. We modified the ASPP module to replace global average pooling with average pooling, with an 8x8 kernel size and a stride of eight. This adaptation allows the model to accommodate SAR images with different dimensions than the ones used during training. We utilized PyTorch's [43] implementation of ResNet and ASPP with the modifications described above to assemble the full model.

The entire model has roughly 1.4 M parameters, 0.7 M in the encoder and the remaining 0.7 in the decoder. We initialize the encoder using ImageNet weights. The decoder is initialized with random weights. The encoder downscales the input image by a factor of eight, and the decoder does not change the height or width of the feature maps. Instead, the model upsamples the decoder output to the input resolution using bilinear interpolation, producing outputs that match the input's spatial dimensions, similar to the strategy used in [44] as implemented in PyTorch. One advantage of performing the pixel classification at a smaller scale is to reduce the chance of misclassifying small artifacts that might be dominated by noise. A caveat of classifying scene at lower resolution and using bilinear interpolation to increase resolution is that small features might be neglected, however, this is less likely to affect model performance due to the polygonal nature of sea ice charts that are used as labels.

B. Training and Testing Methodology

Our training process starts by randomly selecting 20 files to be used for validation. Considering the primary dataset contains 513 scenes, and that we held out 10 scenes to be part of the test set, and another 12 scenes due to limited number of labeled pixels, the train set then consists of 471 scenes. Due to memory limitations and to increase the variability of training samples, we do not use full scenes for training. Instead, we randomly select patches from the 471 scenes. Randomly selecting patches can be interpreted as a data augmentation technique frequently called random-crop augmentation. Hermann et al. [45] found that random-crop augmentation increases texture bias, meaninig models trained with random-crop tend to use texture as an important feature. Texture bias might be an advantage in sea ice characterization as the shape of sea ice chart polygons (not to be confused with floe shape) does not contain meaningful information about its contents. We defined the patch size to be 992 x 992 pixels, equivalent to 79,360 x 79,360 meters considering a pixel size of 80 meters. As described in Section IV-A, the encoder downscales the input features by a factor of eight, meaning the decoder produces outputs of size 124 x 124 pixels that are upsampled back to 992 x 992 using bilinear interpolation. To randomly place the patches across the 471 scenes, our program randomly selects one scene and one patch inside that scene. The program continues to randomly select scenes and patches until it reaches the desired number of patches. We accepted training patches when they contained less than 30% invalid pixels. We considered a pixel invalid if it did not contain labels. Land areas are also marked as invalid, as maps and location of land are known. We defined the 30% threshold to reduce the number of samples containing few labels that would contribute to the loss calculation. In the primary dataset, the invalid values in SAR input features are set to zero, therefore we did not modify them. Loss and metrics ignore the invalid pixels. We randomly selected 4800 patches for each experiment described in this manuscript. Patch extraction took roughly 1 hour using an Intel Xeon CPU 3.7 Ghz and 128 GB RAM. We trained the models using a dual-NVIDIA RTX A5000 Graphics Processing Unit (GPU). One training epoch took roughly 2.2 minutes.

Each trained model can generate outputs for full size Sentinel-1 EW images usually in less than 20 seconds using the CPU described above, or under 5 seconds with one GPU. We calculate loss and metrics for scenes in the validation and test set using the entire scene, which also facilitates the interpretation of validation and test results without requiring patch stitching.

We trained the models using Dice [40] and cross-entropy loss functions to study the effect of loss function choice on uncertainty characterization. We used a batch size of 32 and the AdamW [46] optimizer with default parameters, except for the learning rate. The learning rate started at 1e-4 and was multiplied by 0.1 if the validation loss did not improve for four epochs down to a minimum of 1e-8. The models stopped training when the validation loss did not improve for 10 epochs. The models' weights at the lowest

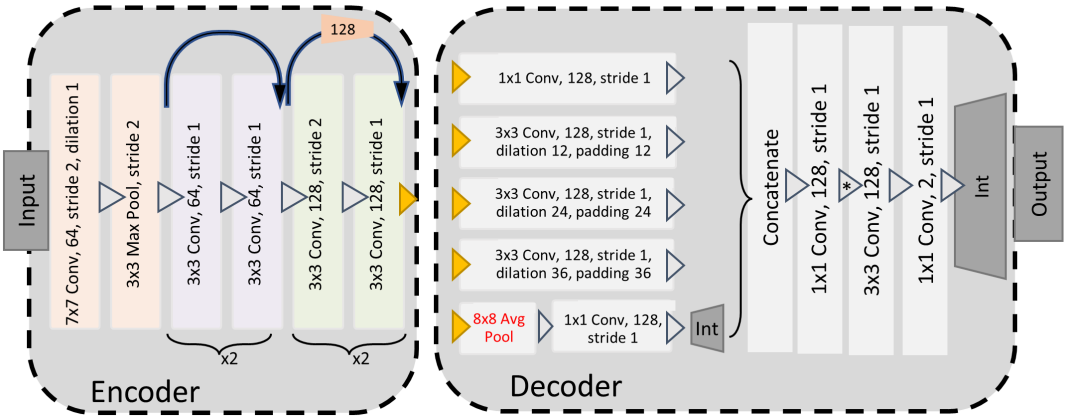


Fig. 3. Model encoder-decoder architecture. Conv stands for convolution, Max Pool for max pooling, Avg Pool for average pooling, Int for bilinear interpolation. The model takes as input arrays with three channels - HH, HV, and incidence angle; and outputs two values representing the occurrence of water or ice. The encoder is based on the ResNet architecture. The orange trapezoids indicate 1x1 convolution. The decoder of the model is adapted from ASPP with the main modification highlighted in red (Avg Pool). The golden triangles indicate how the output of the encoder is used as input for the decoder. Convolution (Conv) layers are followed by batch normalization and activation functions (rectified linear unit for all non-classifier layers) not shown in the figure. The asterisk (*) in c) marks the Dropout location (p = 0.5).

validation loss were saved and later used to generate outputs for validation and test scenes. We evaluated the models' performance using traditional metrics, including F1 and intersection over union (IoU). We used PyTorch [43], PyTorch Lightning [47], and Scikit-Learn [48] as the main libraries to develop our models and analysis. Our code is available at https://github.com/geohai/sea-ice-binary-ai4seaice.

We repeated experiments ten times to verify the models' performance metrics against the stochastic nature of training deep networks, as well as to generate different models for the ensemble. The only difference in each experiment realization is different starting seeds for the pseudo random number generators. Different seeds will select different files for validation as well as different training patches boundaries. We initially experimented with Adam [49], different batch sizes, different learning rates, and we found some improvements using the hyperparameters described in the previous paragraph.

C. Ensemble and Dropout

We use a strategy similar to [35] to generate outputs and estimate model uncertainty. The ensemble output mean probability is simply the mean of the probability output for different realizations. The prediction uncertainty is the Shannon entropy of models' prediction. Shannon entropy (H) of the discrete random variable X is given by:

$$H(X) = -\sum_{i} p(x_i) \log_2 p(x_i) \tag{1}$$

where $p(x_i)$ is the probability of event (x_i). In our setup, $p(x_i)$ is the probability of ice assigned by the model for each pixel output.

Realizations can be generated with two strategies:

1) Enabling dropout during test time [33]: By deactivating different neurons, the output of the model changes slightly even when using the same input. We identify results obtained with this technique as Monte Carlo Dropout (MCD).
2) Ensembling models' predictions [35]: Models trained with the same objective, but with small variations in the values of their parameters, will have different weights and generate different outputs for the same input. As we repeated the experiments changing the sample patches used for training and validation, and the random seed initializing the parameters of the decoder as described in Section IV-B, we have several models with this characteristic. We identify results obtained with this technique as Ensemble (Ens).

Naturally, these strategies are not exclusive and can be used at the same time, which is what we do in some of the experiments reported here. When doing so, we identify those results as MCD+Ens.

V. Results

A. Primary dataset

In this section, we first present evaluation metrics on the primary dataset, and then visually present uncertainty characterization examples for this data which were used in the training of the models.

Fig. 4 shows the evolution of loss during training for all experiments. Some models stopped training with fewer epochs as we used patience to stop training if there was no improvement in the validation loss for 10 consecutive epochs, with a maximum of 50 epochs. The cross-entropy loss is greater than the Dice loss over epoch iterations, which is more a reflection of the scale of the losses than performance differences. The training set loss has small variation across experiments, in contrast to the validation set, which is an indication that some scenes are harder to classify than others. The training and validation loss are at the same order of magnitude. As described before, the validation is composed of 20 scenes randomly selected for each iteration and we conduct

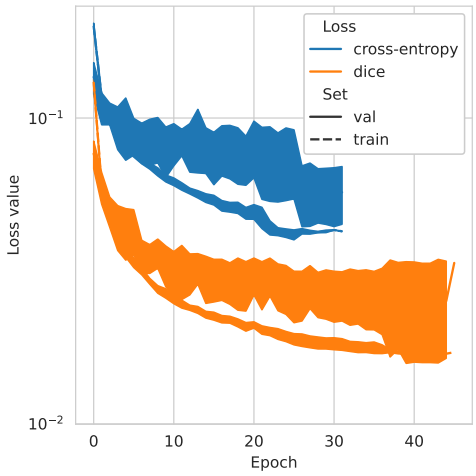


Fig. 4. Training and validation aggregate for all experiments. Colors indicate loss type; line style indicate set. The lines show the median values across 10 different runs, and the 95% confidence intervals are represented with a lighter background band. The models stop training if the validation loss does not decrease in 10 epochs, thus some models train for longer than others and the background aggregate is not displayed if there is only a single value for those epochs.

TABLE I
VALIDATION AND TEST METRICS FOR THE PRIMARY DATASET. ROWS HIGHLIGHTED IN BOLD CORRESPOND TO MODELS USED IN FIGS. 5 TO 9.

	wIoU		wF1	
Exp.	**Validation**	**Test**	**Validation**	**Test**
ce-1	0.97 ± 0.04	0.96 ± 0.04	0.98 ± 0.02	0.98 ± 0.02
ce-2	0.95 ± 0.05	0.96 ± 0.03	0.97 ± 0.03	0.98 ± 0.02
ce-3	0.95 ± 0.04	0.96 ± 0.04	0.97 ± 0.02	0.98 ± 0.02
ce-4	0.94 ± 0.07	0.96 ± 0.03	0.97 ± 0.04	0.98 ± 0.02
ce-5	0.95 ± 0.05	0.96 ± 0.04	0.97 ± 0.03	0.98 ± 0.02
ce-6	0.97 ± 0.03	0.96 ± 0.04	0.98 ± 0.01	0.98 ± 0.02
ce-7	**0.94 ± 0.06**	**0.96 ± 0.04**	**0.97 ± 0.03**	**0.98 ± 0.02**
ce-8	0.94 ± 0.04	0.96 ± 0.04	0.97 ± 0.02	0.98 ± 0.02
ce-9	0.93 ± 0.09	0.96 ± 0.03	0.96 ± 0.05	0.98 ± 0.02
ce-10	0.96 ± 0.05	0.96 ± 0.04	0.97 ± 0.03	0.98 ± 0.02
dice-1	0.97 ± 0.03	0.96 ± 0.04	0.98 ± 0.02	0.98 ± 0.02
dice-2	0.95 ± 0.05	0.96 ± 0.04	0.97 ± 0.03	0.98 ± 0.02
dice-3	0.95 ± 0.04	0.96 ± 0.03	0.97 ± 0.02	0.98 ± 0.02
dice-4	0.94 ± 0.07	0.96 ± 0.03	0.97 ± 0.04	0.98 ± 0.02
dice-5	**0.95 ± 0.05**	**0.96 ± 0.04**	**0.98 ± 0.03**	**0.98 ± 0.02**
dice-6	0.97 ± 0.03	0.96 ± 0.04	0.98 ± 0.01	0.98 ± 0.02
dice-7	0.94 ± 0.07	0.96 ± 0.04	0.97 ± 0.04	0.98 ± 0.02
dice-8	0.94 ± 0.05	0.96 ± 0.04	0.97 ± 0.03	0.98 ± 0.02
dice-9	0.94 ± 0.07	0.96 ± 0.03	0.97 ± 0.04	0.98 ± 0.02
dice-10	0.96 ± 0.05	0.96 ± 0.04	0.98 ± 0.03	0.98 ± 0.02

the analysis using full scenes. As we are more interested in understanding how the model performs with data not used during training, the remainder of this paper focus on either the validation or, most frequently, test results.

Table I shows metric summary for several experiments. The table indicates the loss used (ce for cross-entropy) and the metrics obtained in the validation and test sets. Table I also indicates the experiment number, weighted Intersection over Union (wIoU), and weighted F1 (wF1) metrics. The metrics are weighted based on the number of water and ice pixels. The metrics show the average computed across the scenes in each one of the sets, as well as the standard variation. Although the validation scenes are different for each experiment generating some variation in performance for different experiments, there is less oscillation on the metrics for the test set within the same loss. In fact, test wF1 reported with two decimal points is the same for all experiments. These results indicate that the models are stable and, although possibly generating slightly different results, can achieve wIoU and wF1 above 0.95 for the primary dataset.

The test scene with consistently higher metrics was 20201112T080407_dmi_prep. The wF1 average across all 20 experiments was 0.995, wIoU average was 0.989, and average accuracy was 0.995. Fig. 5 shows one example of the classification for scene 20201112T080407_dmi_prep, where the model has high performance in classification, and the corresponding Shannon entropy, simply entropy hereinafter, as characterization of uncertainty. The thin line observable errors are at the ice-water border and are likely associated with the lower encoder resolution as described in Section IV-A. Although still small, some of the larger error regions might be caused by the fact that primary dataset polygons (used in training) are not as detailed as the secondary dataset shown in the figure, leading the model to miss small spatial details.

To stay consistent in visualization, throughout the remainder of the manuscript, we clip ice probability in the range [0.2, 1.0] and entropy in the range [0.29, 1.0]. A probability of 0.95 corresponds to an entropy value of 0.29 (eq. 1). Maintaining the same scale throughout images facilitates the comparison across results in this study. The scene with consistently lower metrics was 20210715T211029_dmi_prep, a scene acquired during summer in the Northern Hemisphere. Sea ice characterization based on SAR imagery is known to be more challenging during summer due to surface melt. Fig. 6 shows the results for the worst performing model for that scene. Comparing results in Fig. 5 and Fig. 6,we see that the open water conditions are different, and surface melt is occurring on sea ice. The model misclassifies a large region of sea ice as water on the left-center side of the figure, for example. The input to the model in Fig. 6 highlights the importance of sea ice analysts' interpretation skills, and the subsequent quality control necessary on deep learning output. There is noise contaminating the HV input in Fig 6b, especially strong vertical banding effect. Fig. 7 shows the confusion matrix for results in Fig. 6. The main error is the misclassification of ice as water.

Cases like the one in Fig. 6 are not uncommon. Machine learning models, or any type of model, can generate outputs that are erroneous interpretation of the reality. It is challenging and counterproductive for users to quality-control all machine learning outputs. This highlights the value that prediction uncertainty quantification adds to the decision-making process.

Fig. 8 shows the Ens mean ice probability and prediction uncertainty assigned by the 10 models trained with Dice loss for the lowest scoring 20210715T211029_dmi_prep scene. Results show a very high ice probability of ice within the ice region, and ice probability close to zero for the water region, even for areas misclassified in Fig. 6. The differences between models' outputs is more evident in the prediction uncertainty

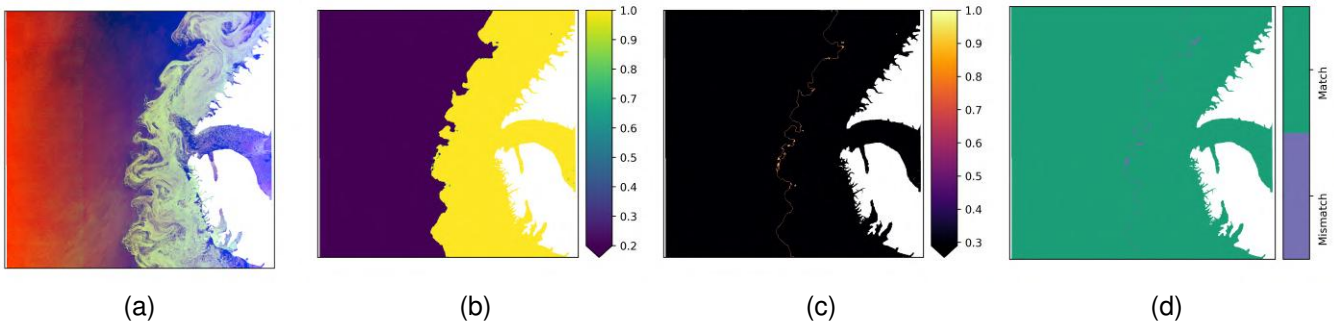


Fig. 5. Test set example (20201112T080407_dmi_prep). (a) Shows the RGB composition of the inputs to the model, HH, HV, and incidence angle. (b) Shows ice probability assigned by the model (dice-5). (c) Shows the entropy associated with the probability assigned by the model. (d) Shows the mismatch between the model output (dice-5) and the labels in the dataset. This corresponds to accuracy and wF1 of 0.996, and an wIoU of 0.992

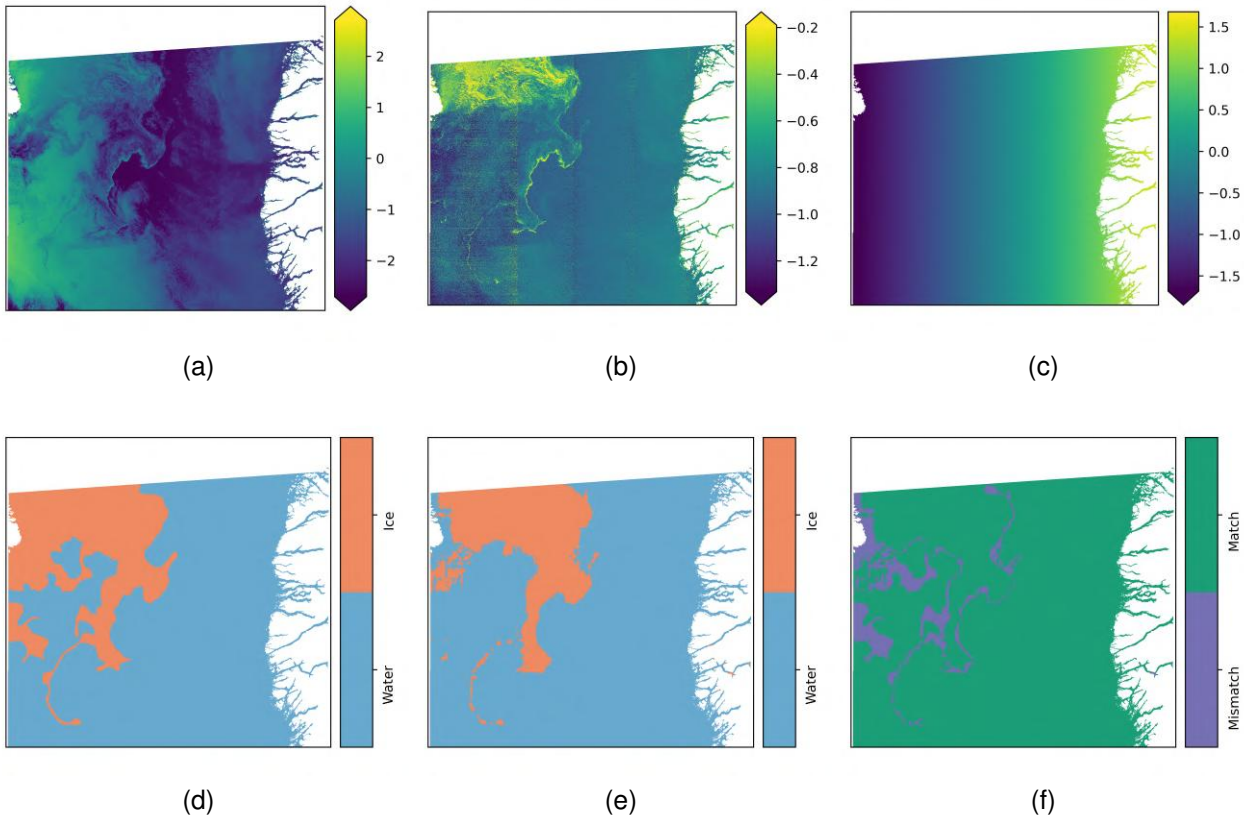


Fig. 6. Lowest scoring test set example (20210715T211029_dmi_prep) in the primary dataset. (a) HH, (b) HV, and (c) incidence angle are the input to the model. (d) The primary dataset labels and (e) Model classification (ce-7) share the same colorbar. (f) Shows the mismatch between (d) and (e). This corresponds to accuracy and wF1 of 0.92, and an wIoU of 0.86.

map in Fig. 8b. The prediction uncertainty image highlights a thin line around the ice boundaries and weakly identify larger regions of disagreement between model outputs. Results in Fig. 9 are analogous as the one in Fig. 8, but now for the cross-entropy loss model ensemble. The misclassified areas are more evident as uncertain areas in Fig. 9 with cross-entropy loss, which might be preferable when identifying challenging areas, where human quality-control may be necessary to override misclassifications.

B. Secondary dataset

In this section, we present the evalation metrics on the secondary dataset which was used only for testing (and not training), and then analyze uncertainty characterization using different strategies. We will first need to identify areas of strength and weakness in classification to then focus on uncertainty characterization in those areas, thus our approach in carefully establishing error metrics and visual analyses. The purpose of this test on a secondary dataset is to investigate the generalizability of the models and uncertainty characteriza-

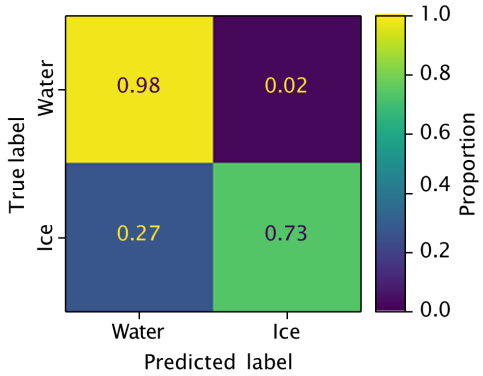
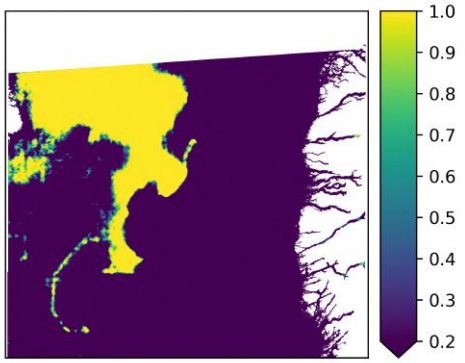
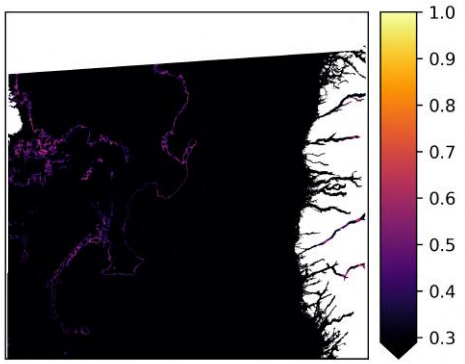


Fig. 7. Confusion matrix for results in Fig. 6. The main error of the model is to misclassify ice as water. The confusion matrix is normalized per row.



(a)

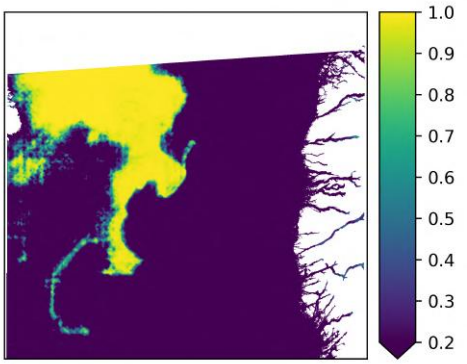


(b)

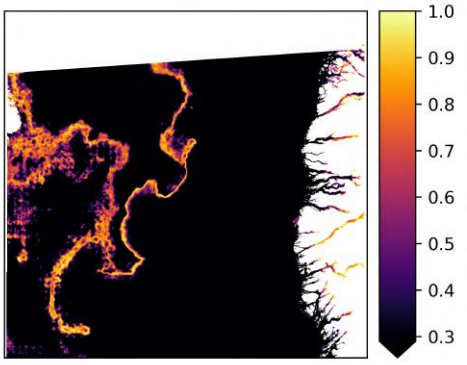
Fig. 8. (a) Ens mean ice probability output from 10 models trained with Dice loss. (b) Ens prediction uncertainty computed with the same 10 outputs.

tion under circumstances similar to operational environments, where a single dataset is initially used for training the models, which are then used for classifying unseen imagery (such as in our secondary dataset) for operational use.

We experimented with MCD+Ens for the secondary dataset analysis. To do so, we process each scene 10 times with the same model with dropout enabled, and we repeat the same strategy for each one of the 10 models trained previously. As a result, the output for one scene is an aggregation of 100 estimates. As mentioned before, the secondary dataset was



(a)



(b)

Fig. 9. (a) Ens mean ice probability output from 10 models trained with cross-entropy loss. (b) Ens prediction uncertainty computed with the same 10 outputs.

TABLE II
TEST METRICS FOR THE SECONDARY DATASET USING MCD+ENS

Loss	Accuracy	wIoU	wF1
Cross-entropy	0.92 ± 0.06	0.87 ± 0.07	0.93 ± 0.05
Dice	0.93 ± 0.05	0.88 ± 0.07	0.93 ± 0.04

never used in training and underwent a different pre-processing stage, and were labeled by different sea ice analysts. Table II shows a summary of the metrics for the secondary dataset using MCD+Ens. The performance metrics are first computed independently for each scene, then aggregated for the 12 scenes in the dataset. Evaluation metrics show a performance decrease in comparison to the metrics computed for the primary dataset. Dice results are marginally better than cross-entropy results.

Fig. 10 shows the classification errors maps for each one of the 12 scenes in the secondary dataset using models trained with cross-entropy loss. Results in Fig. 10 are provided as an overview and more details are provided for specific regions and in Section V-C. In the interest of avoiding too many figures and still showing most of the results, we show only the error maps and not the input and labels for each scene. Interested readers can find the list of Sentinel-1 scenes, as well as their associated labels in [12]. Most scenes show a thin line of error at the ice-water boundary, similar to what is observed for the primary dataset. However, results in Fig. 10 show lower

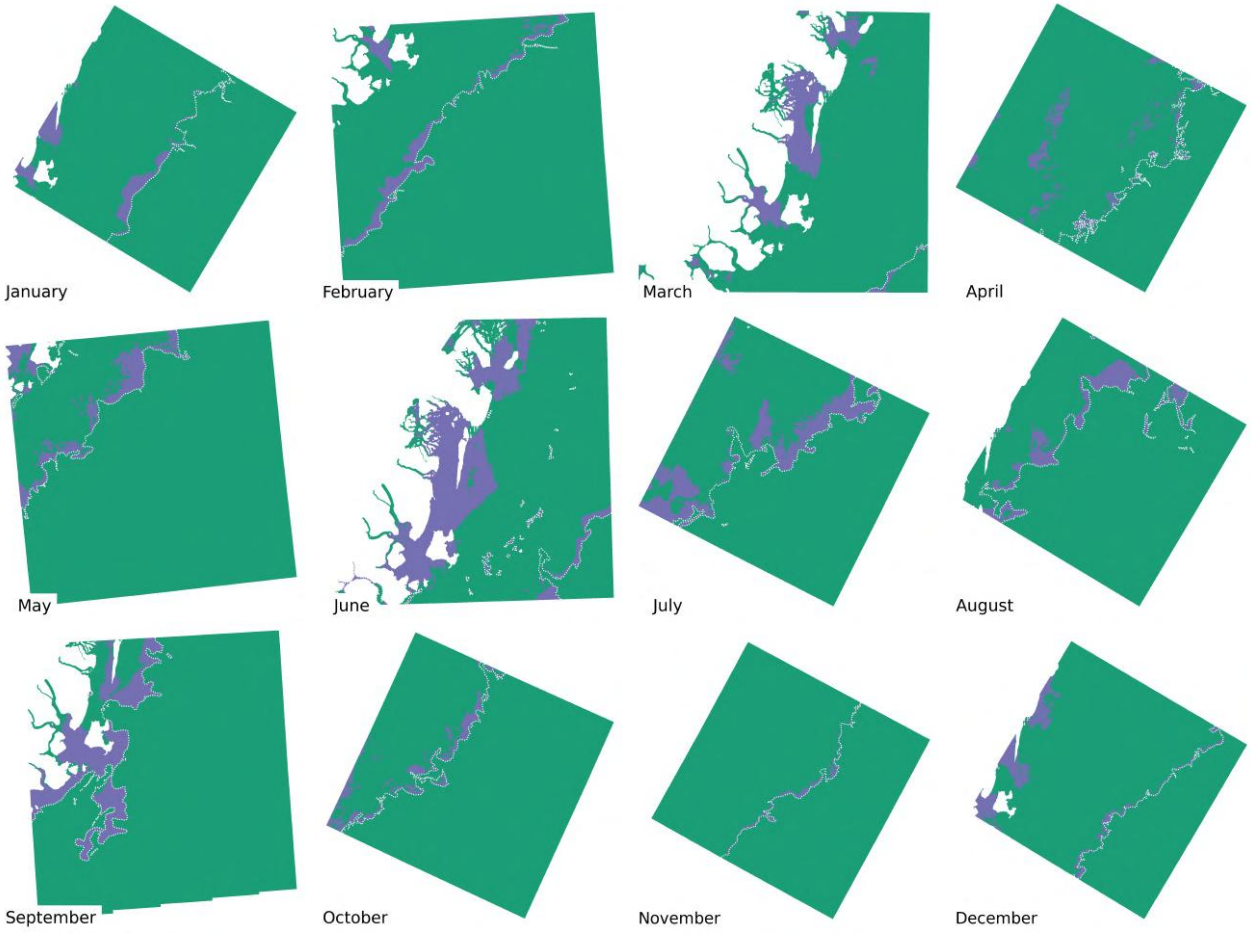


Fig. 10. Error maps for the twelve images in the secondary dataset. A dotted line shows the label limits for water and ice as provided in the secondary dataset. All scenes have most of the labeled ice concentrated to the West or Northwest of the dotted line (North is up). All panels use the same colormap: green for matching and purple for mismatching values. Projection details are in Appendix B. All scenes have roughly the same physical size (400 x 400 km), but some scenes appear larger than others here due to projection. Each panel shows the image for each month starting from January on the top left. September, May, and August have rectangles locating Figures 12, 14 and 15.

performance for June, July, August, and September, associated with larger regions of misclassifications. The scene with best performance is November, with accuracy and wF1 of 0.99 and wIoU of 0.98. Fig. 11 highlights the results obtained for January, a scene with accuracy and wF1 of 0.96 and wIoU of 0.92. The input to the model suffers from strong TOPSAR banding noise clearly defined by the linear features over open water on the SW of the image. As observed in the primary dataset, the entropy highlights the region where the ice boundary predictions end. However, as the white line indicates, the models tend to miss some regions where new ice is forming and appears darker in the Sentinel-1 scene when compared to the remaining of the ice in that image. The models also produce uncertainty regions for areas close to land where they struggle to correctly predict ice that shows as blue in the SAR image.

In order to set the stage for uncertainty characterization, we need to take a closer look at areas of high and low performance. Table III shows scene-by-scene precision and recall for ice for each sample in the secondary dataset. The mean ice precision is 0.99 for Dice loss and 1.00 for cross-entropy. This indicates that pixels indicated as ice by the models indeed contain ice. Lower values of ice recall show that the models are not able to classify all ice regions. The lowest ice recall values are for September, a scene covered mostly by water and where the interpretation is challenging. Fig. 12 shows an inset for September that can be located back in Fig. 10. The MCD+Ens entropy in 12c captures the part of the ice that is most visible in Fig. 12a as an uncertain region, but not the entirety of fist year ice interpreted in Fig. 12b.

C. Ensemble or Dropout for prediction uncertainty estimation

This section provides a comparison of the two methods we used for prediction uncertainty estimation. For the MCD predictions, we arbitrarily selected the seed #1 models trained with cross-entropy and Dice losses. To generate different outputs, we enabled dropout during test time and generated ten realizations with the selected models. We generated Ens results by simply using all 10 trained models for each loss to obtain an output. Fig. 13 shows a comparison of the prediction uncertainty using different strategies for April in the secondary dataset. For both strategies, the uncertainty covers most, if not all, of the ice-water boundary. Although this is

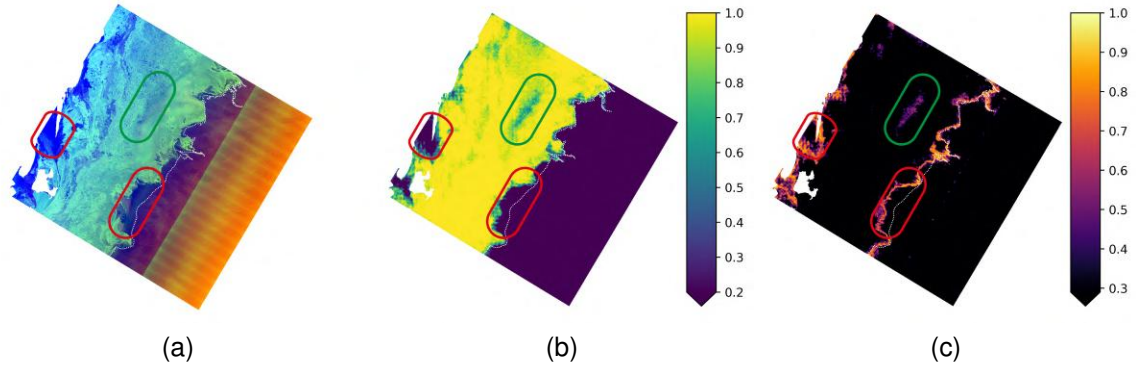


Fig. 11. Secondary dataset, January. (a) RGB composition of the inputs to the model, HH, HV, and incidence angle. (b) Mean ice probability from the ensemble of 10 models trained with cross-entropy, 10 realizations for each model with dropout enabled (MCD+Ens). (c) Entropy associated with the probability in (b). The red polygons indicate misclassifications in regions of low uncertainty; the green polygon points to a region of high uncertainty that was correctly classified.

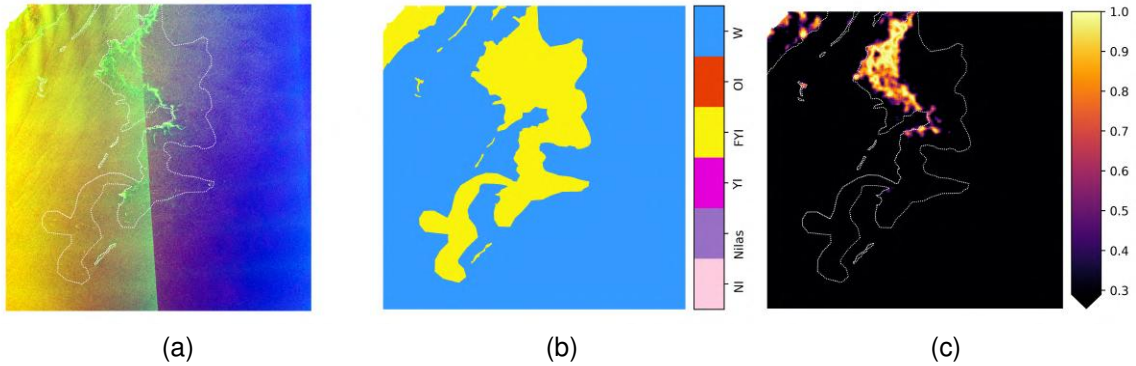


Fig. 12. Secondary dataset, September inset example. (a) RGB composition of the inputs to the model, HH, HV, and incidence angle. (b) Ice chart from the secondary dataset colored by the oldest ice type: New Ice (NI), Nilas, Young Ice (YI), First Year Ice (FYI), Old Ice (OI) and Water (W). (c) MCD+Ens Entropy. The panels in this figure represent a square region with size 150km^2

TABLE III
SCENE-BY-SCENE PRECISION AND RECALL FOR ICE IN THE SECONDARY DATASET USING MCD+ENS

	Cross-entropy		Dice	
Scene	**Precision**	**Recall**	**Precision**	**Recall**
Jan	1.00	0.92	1.00	0.91
Feb	1.00	0.87	1.00	0.88
Mar	0.99	0.89	0.99	0.88
Apr	0.99	0.94	0.98	0.94
May	1.00	0.75	1.00	0.78
Jun	0.99	0.75	0.99	0.79
Jul	1.00	0.74	1.00	0.77
Aug	1.00	0.77	0.99	0.80
Sep	0.99	0.34	0.99	0.35
Oct	1.00	0.88	0.99	0.92
Nov	1.00	0.98	1.00	0.99
Dec	1.00	0.92	1.00	0.94
Mean	**1.00**	**0.81**	**0.99**	**0.83**
Std	**0.00**	**0.17**	**0.01**	**0.17**

expected and unlikely to prompt even more attention from sea ice analysts that would likely quality control the ice-water boundary regardless of the results of the model, the fact that most of the uncertainty is concentrated on icy regions is a potential advantage to draw attention to those ice-infested areas. There is no uncertainty associated with open water regions, even though the input in Fig. 13a shows banding noise over open water. Both MCD entropy (Fig. 13c) and Ens entropy (Fig. 13d) identify with higher uncertainty the first year ice region (Fig. 13b) misclassified in Fig. 10. Although the MCD uncertainty in Fig. 13c shows brighter values, they are more concentrated and not as diffuse as results in Fig. 13d, where regions too large are identified as uncertain.

The fact that our models tend to generate higher uncertainty for ice-infested regions is a potential advantage in leading the attention of sea ice analysts for quality control. Although sometimes the banding noise in Sentinel-1 images can create some uncertainty over open water, as barely visible in Fig. 11c, in most cases the uncertainty over water is very low as discussed in the previous paragraph. Fig. 14 further highlights the higher MCD uncertainty over ice when compared to water regions. The location of the data in Fig. 14 is highlighted in Fig. 10. The Sentinel-1 banding noise is noticeable in Fig. 14a, specially over water. Although it might be hard to label the boundary exactly like the interpretation in Fig. 14b, it is relatively easy to see the ice as a brighter region than the surrounding water. The models, however, misclassify part of the ice as water, as Fig. 14d shows. The higher MCD uncertainty in Fig. 14c provides a mean to quickly identify regions that were potentially incorrectly classified.

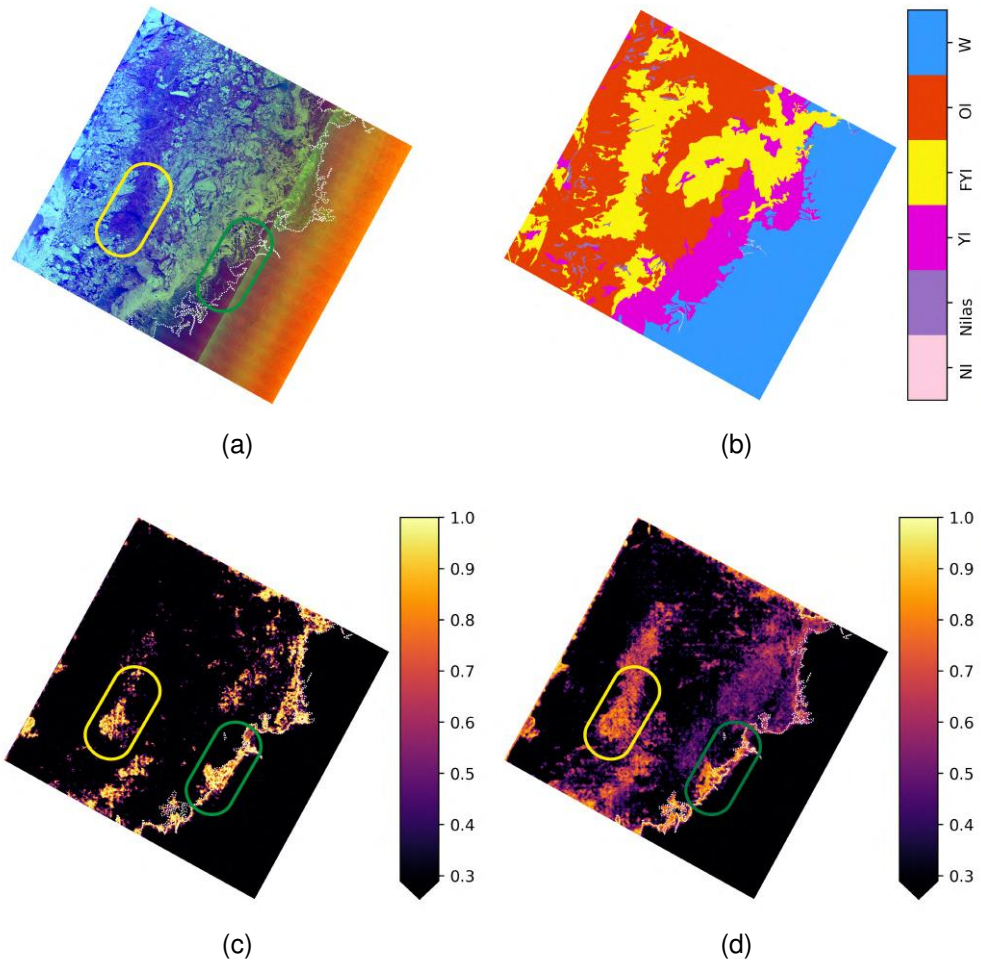


Fig. 13. Secondary dataset, April example. (a) RGB composition of the input to the model: HH, HV, and incidence angle. (b) Ice chart from the secondary dataset colored by the oldest ice type: New Ice (NI), Nilas, Young Ice (YI), First Year Ice (FYI), Old Ice (OI) and Water (W). (c) MCD entropy. (d) Ens entropy. The polygons in this figure point to two regions where the uncertainty is helpful for a sea ice analyst. The green polygon indicates a region of high uncertainty on which the model assigned the same class as the labels. The yellow polygon indicates a region where the model assigned high uncertainty and assigned a class different than the labels.

In general, Ens generates more diffuse uncertainty boundaries than MCD, and models trained with cross-entropy tend to generate more diffused uncertainty than models trained with Dice. Fig. 15, whose location can be traced back to Fig. 10, provides an uncertainty comparison across all these. Fig. 15g shows that MCD entropy for models trained with Dice loss have a much thinner uncertainty boundary than Fig. 15e for the MCD entropy for models trained with cross-entropy loss. In fact, some of the brighter and isolated floes in Fig. 15a seem to be highlighted in Fig. 15g. As Ens tend to have more diffuse uncertainty than MCD, the results obtained for MCD+Ens are very similar to those obtained by Ens only and not really noticeable at this scale. Thus, Figs. 15d and 15f are very similar, just like Figs. 15h and 15i. When comparing MCD entropy with Ens entropy, we note that uncertainty brightness tends to be higher for MCD. In fact, results in Figs. 15h and 15i are dim. This is a reflection of the overconfidence of ensemble models trained with Dice loss and the entropy computing strategy. For MCD, the prediction variation is smaller and more localized, therefore the uncertainty from the prediction realizations align with each other. Therefore, when we compute the average entropy across realizations, the same pixels will have large entropy for different realizations, producing a larger final entropy average. For Ens, the uncertainty regions do not align as well as MCD, bringing the average down as pixels with large entropy in one realization might not have large entropy in another realization. The generally lower confidence of cross-entropy models better highlights regions of predictions uncertainty. Although still unable to fully delineate the same ice-water boundary that was interpreted by a sea ice analyst, the entropy of the cross-entropy models in Figs. 15e, 15d, and 15f better capture that challenging region.

Fig. 16 shows a histogram for the probability assigned by different models for four scenes in the secondary dataset. Results show that, in general, cross-entropy generates results that are better calibrated when compared to Dice as the probabilities bins are slightly more balanced, i.e., more pixels were predicted with intermediate probabilities between 0.1 to 0.9. and fewer pixels were predicted with probabilities above 0.95. In general, Ens predictions also appear more calibrated.

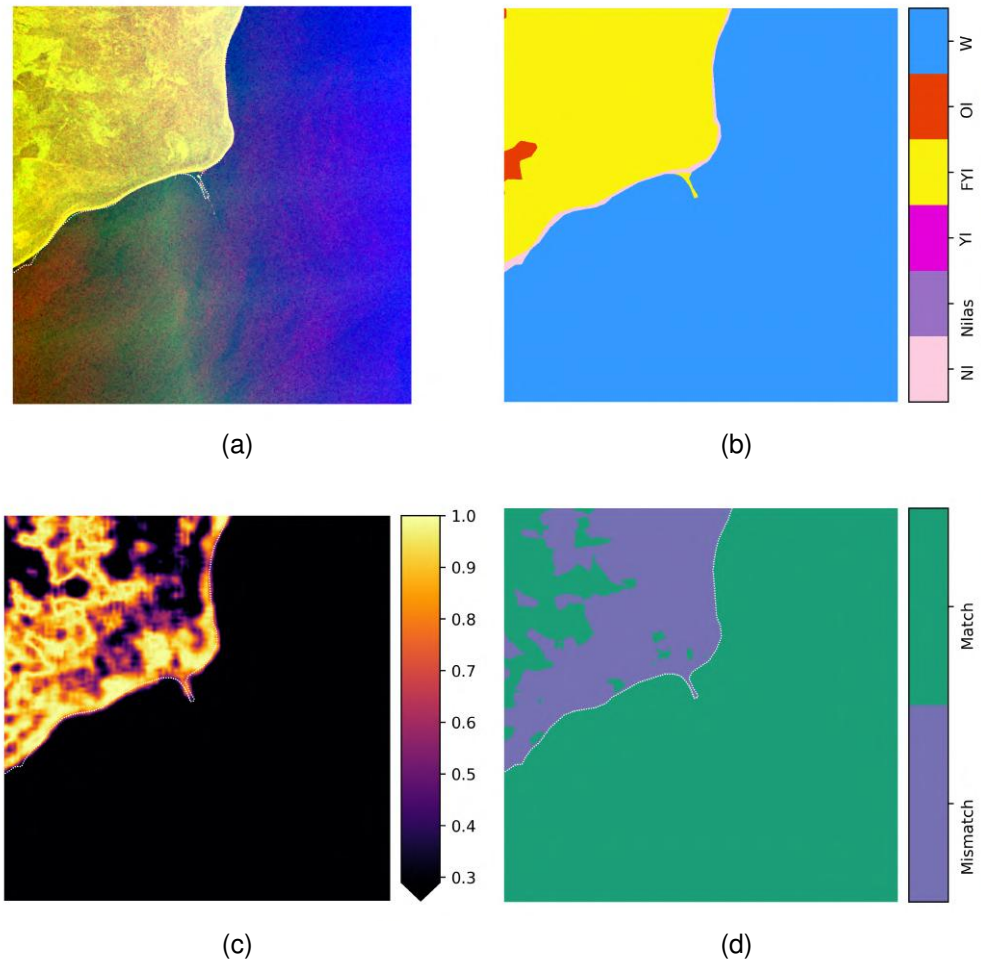


Fig. 14. Secondary dataset, May inset example. (a) RGB composition of the input to the model: HH, HV, and incidence angle. (b) Ice chart from the secondary dataset colored by the oldest ice type: New Ice (NI), Nilas, Young Ice (YI), First Year Ice (FYI), Old Ice (OI) and Water (W). (c) MCD entropy. (d) MCD error. The panels in this figure represent a square region with size 80km^2.

However, this is not always the case. In Fig. 16a, Ens predicted fewer pixels with intermediate probabilities between 0.1 to 0.7 than its single prediction counterpart.

Besides providing slightly better calibration, cross-entropy uncertainty also appears to have an advantage as a means of identifying regions that require further quality control. Fig. 17 shows mean accuracy of pixels below a certain entropy threshold for four scenes in the secondary dataset and indicates that the entropy of the models' predictions is a viable way to obtain some information on poor quality results. Results in Fig. 17 show that the mean accuracy decreases as entropy increases, as expected. However, it also becomes evident that predictions generated with models trained with dice loss are more significantly affected by overconfidence. While the accuracy of the predictions of models trained with cross-entropy decreases as entropy increases, dice mean accuracy falls below 0.3 for all scenes even at the lowest entropy threshold of 0.01 (roughly equivalent to a probability of 0.999). This indicates that for models trained with dice, most pixels are correctly classified as ice only when the probability assigned to ice is ≈ 1. Ideally the accuracy would drop less aggressively for different thresholds. Although cross-entropy accuracy falls faster than ideally desired, its behavior is again slightly better when compared to dice. Results in Fig. 17 also indicate that Ens shows marginally higher accuracy at the same entropy threshold than MCD.

VI. Discussion

Our results indicate that MCD and Ens generated entropy can be used to highlight areas of uncertainty in deep learning predictions using C-band SAR. However, MCD has some advantages over Ens, specially when we consider the results for the secondary dataset. MCD entropy can be generated with just one model while keeping dropout active, whereas for the ensemble strategy, several models need to be trained. Training multiple models may not be practical in operational scenarios where time and computational resources are limited. Additionally, MCD generates more focused areas of uncertainty, as opposed to the diffused areas of uncertainty when multiple models are trained in an ensemble. We clarify that in the experiments presented here, the ensemble models have the same architecture and were trained with the same hyperparameters. Their differences are stochastic, in the starting seeds as well as in the samples selected for training and validation. The

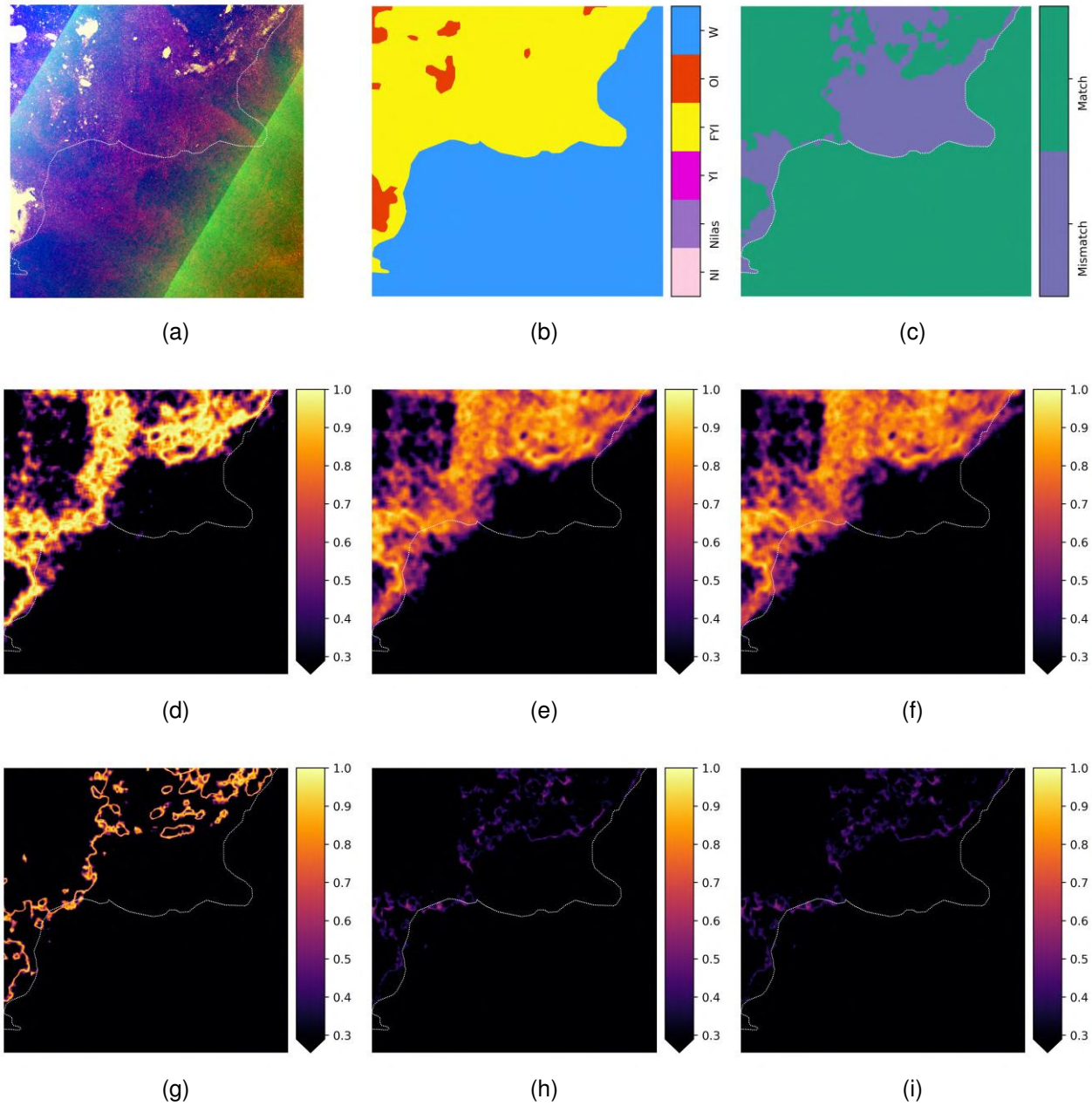


Fig. 15. The effect of loss function, MCD and Ens. for uncertainty visualization. The selected sub-image is from the Secondary dataset, August inset example. (a) RGB composition of Sentinel-1 HH, HV, and incidence angle for the Aug inset in Fig. 10. (b) Ice type labels from the secondary dataset (a). (c) Misclassification from cross-entropy MCD+Ens. (d) MCD entropy for models trained with cross-entropy. (e) Ens entropy models trained with cross-entropy. (f) MCD+Ens entropy for models trained with cross-entropy. (g), (h), and (i) are analogous for Dice. The white dashed line shows the ice boundary from the labels in (b). The panels in this figure represent a square region with size 100km^2

more focused areas of uncertainty might be advantageous in directing the human user attention on those challenging areas.

Our MCD strategy is similar to what Asadi et al. [28] used to investigate how weights' uncertainty affects the predictions of the models. However, based on visual inspection, Asadi et al.'s [28] uncertainty maps appears more diffuse than our results. We believe our uncertainty maps are not quite comparable, as our study is distinct from Asadi et al.'s [28] analysis in two main points that might explain the difference in results. First, the model architecture we use is based on CNN whereas Asadi et al. [28] uses multi-layer perceptron neural networks. Second, the dataset and in our study are from a wider range of polygons.

Throughout our experimentation, we observed that dice uncertainty was generally more localized than cross-entropy uncertainty. This aligns with what was observed in Mertash et al. [39] that models trained with dice loss tend to generate outputs with higher confidence, i.e., not well-calibrated. In [39], using ensemble was useful to improve the calibration of models trained with dice loss, however we did not see a significant improvement in calibration with ensembling in our experiments, as Fig. 16 shows.

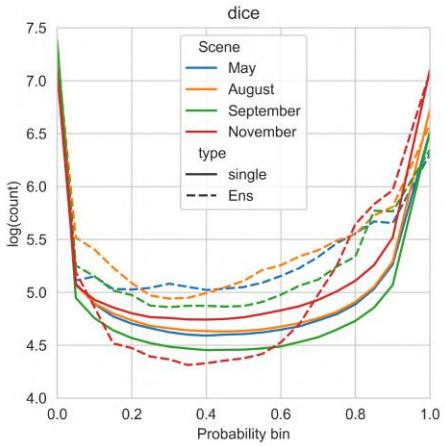


Fig. 16. Probability histogram assigned for selected scenes of the secondary dataset. The y-axis shows the log_{10} for the count of pixels that were assigned the probability in the x-axis. We computed the histogram considering 20 bins. (a) Probability histogram for results obtained with Dice loss. (b) Analogous for cross-entropy loss.

While assessing uncertainty is useful in the majority of cases, it is not full proof when models exhibit unwarranted confidence while making incorrect predictions. This behavior is dependent on several factors, including the similarity of latent feature distributions between unseen images and the training dataset, and similarity of pre-processing algorithms. In Fig. 12, for example, banding noise intersects with areas that are inaccurately classified. It is worth reiterating that the secondary dataset underwent different pre-processing steps, including a different thermal processing algorithm to correct for banding noise and the way intensities are balanced across bands. Although our models displayed higher uncertainty for the region in the Sentinel-1 image with higher intensity labeled as first-year ice, the surrounding area seems to exhibit less roughness texture, and the models classified as water what was labeled as ice. This absence of roughness texture in C-band imagery likely contributed to our models' challenges in accurately categorizing the region illustrated in Fig. 14, although we expected that it would be an easier region for models' prediction due to the higher backscatter intensity. Ultimately, we acknowledge that the models are not perfect, and may generate wrong predictions with low uncertainty.

It is worth reiterating the value of using entropy for characterizing uncertainty. In classification, a thresholding of model probability outputs is used to determine the assigned class. Therefore, a pixel with ice probability of 0.501 for instance, will be assigned the same class as a pixel with ice probability of 0.999. Entropy effectively highlights such areas of lower probability and stretches the range of uncertain values in the uncertainty map. Although our experimentation was limited to binary classification of ice or water, entropy measurements can also facilitate the interpretation of the results of other multiclass sea ice mapping objectives, such as sea ice concentration or stage of development. Frequently, multiclass models generate a probability for each one of the classes. Sea ice analysts could investigate the probability assigned to each class, but summarizing the information onto a single entropy map could facilitate the interpretation of machine learning models and associated uncertainty. With a single map to quality control, sea ice analysts could quickly identify areas on which the model assigned a class with lower probability.

Our results indicate that most areas marked as uncertain in the deep learning output on SAR are over ice-infested waters. However, not all areas that are misclassified are assigned high uncertainty. For instance, the dark smooth ice in January (secondary dataset, Fig 11c) at the ice edge is not marked as having high uncertainty, yet is misclassified. While this may indicate a limitation of our approach, it is worth noting that the secondary dataset labels are annotated not just using Sentinel-1 SAR images, and rather, other data sources are consulted, including optical and infrared remote sensing, and meteorological data. The sea ice analysts who created the ice chart labels especially relied on this non-SAR information for areas where the interpretation is challenging.

The presence of TOPSAR banding noise does not appear to have any noticeable impact on the classification results or the uncertainty values, which is an encouraging observation. Even in the September test image from the secondary dataset where banding noise is clearly visible, the prediction results remain unaffected. Furthermore, there are no distinct regions with high entropy values in the results.

Lastly, our cross-dataset evaluation provides a more reliable indication of model performance if it were to be deployed. While we observe a decrease in performance from primary to secondary dataset, the performance reduction is marginal. The secondary dataset underwent different pre-processing steps, including a different denoising algorithm. Additionally, some apparent errors are caused by the buffer created by the sea ice analyst, where the ice edge is drawn towards open water and not exactly aligned with the ice edge as observed in the SAR image, especially in August and September images. Such buffers are usually created in ice charts as a safety measure to create a safety buffer for marine navigation.

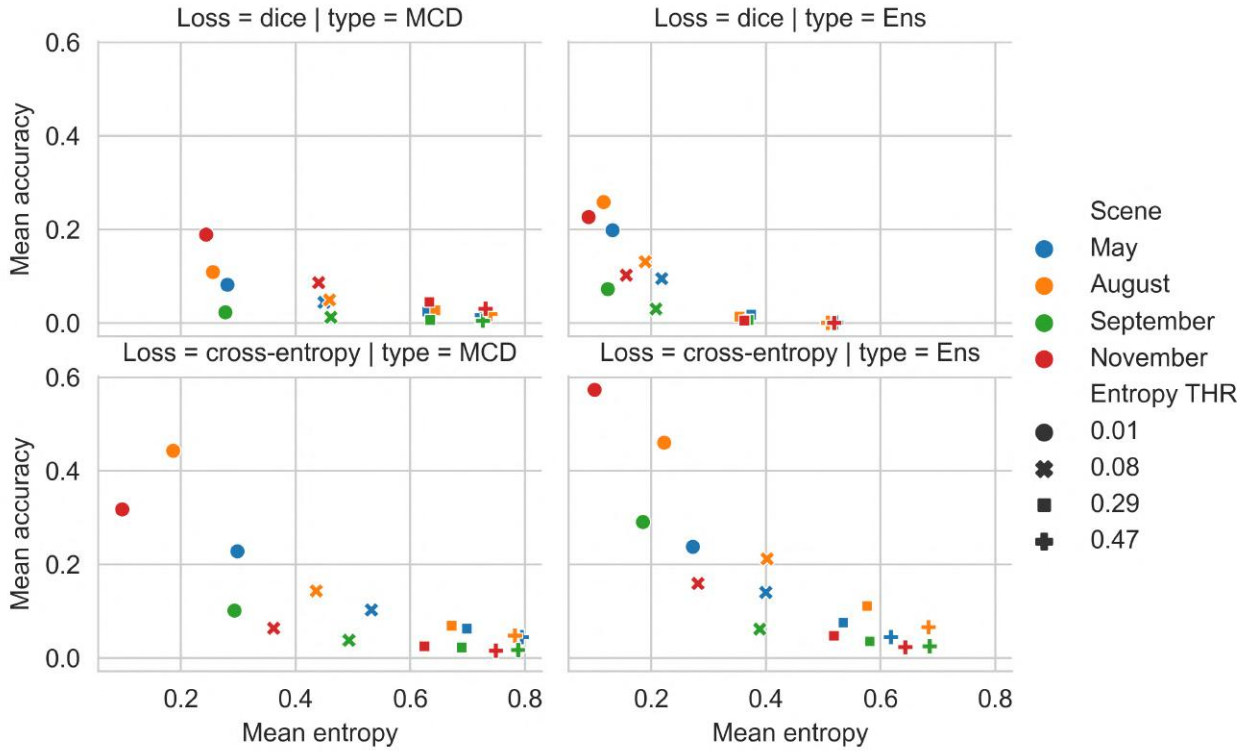


Fig. 17. Mean entropy vs mean accuracy computed at different thresholds for selected scenes of the secondary dataset. THR stands for threshold. Entropy values of 0.01, 0.08, 0.29, and 0.47 correspond to probability values for binary classification of approximately 0.999, 0.99, 0.95, and 0.90, respectively.

VII. Conclusions

We presented and compared two different methods and their combination for characterizing uncertainty in sea ice segmentation using SAR data, and analyzed the effect of loss function choice on the resulting uncertainty. We discussed how models' overconfidence and prediction uncertainty may affect automated products, and how the use of our uncertainty characterization methods can provide means for quality control of the final products. We observed slightly higher test metric performance from models trained with Dice when compared to cross-entropy loss. However, Dice metrics improvements come at the cost of overconfident predictions resulting in smaller and dimmer uncertainty regions for misclassified areas. More pronounced uncertainty indicators might be valuable for highlighting areas that require more attention from the sea ice analysts when doing quality control of output generated by automated models.

We observed that MCD generates concentrated uncertainty with higher values for misclassified areas, as opposed to ensemble methods where uncertainty values are more diffused, thus potentially less helpful to the human analyst. MCD uncertainty estimation is also more computationally-efficient, not requiring multiple trained models. To further understand how a trained model performance varies when used on a dataset with different characteristics, we trained models using the AI4Arctic Sea Ice Challenge, and tested the models on the ExtremeEarth v2, two datasets with different properties and labeling practice. Although there is a decrease in prediction performance when we test the models on ExtremeEarth v2, the results are generally acceptable, with wF1 higher than 90%. We also observe that while summary performance metrics such as accuracy and F1 might not always highlight challenging scenes; our uncertainty characterization methods help in the identification of regions in need of quality control. In general, we did not observe strong evidence of misclassification or increased prediction uncertainty associated with Sentinel-1 denoising techniques, projection, or pre-processing, which is a positive aspect considering national ice centers might have different pre-processing practices.

VIII. Acknowledgment

We would like to thank Behzad Vahedi for providing the pre-processed Sentinel-1 images for the ExtremeEarth dataset. We also thank the anonymous reviewers that helped us clarify and improve our manuscript.

A. List of files from AI4Arctic Sea Ice Challenge Dataset never used in training

Note: Participants of the challenge identified that scene 20200625T081801_dmi_prep was incorrectly labeled. This information was brought to our attention during the final stages of analysis of this paper. Therefore, our models were trained with one incorrectly labeled scene.

Test selection: Images with stage of development labels for 70% to 80% of pixels considering images of size 5000 x 5000 pixels:

20180903T155253_cis_prep.nc

20191016T155300_cis_prep.nc
20201104T171455_dmi_prep.nc
20200124T102732_cis_prep.nc
20201112T080407_dmi_prep.nc
20210208T081803_dmi_prep.nc
20211015T120121_cis_prep.nc
20210429T080105_dmi_prep.nc
20210517T080142_dmi_prep.nc
20210715T211029_dmi_prep.nc

Discarded samples: Files containing ice concentration labels for less than 10% of the pixels:

20190929T140604_cis_prep.nc
20191028T132359_cis_prep.nc
20180903T155153_cis_prep.nc
20190924T144554_cis_prep.nc
20190503T104149_cis_prep.nc
20210523T121414_cis_prep.nc
20191028T132259_cis_prep.nc
20190929T140504_cis_prep.nc
20201016T082722_dmi_prep.nc
20200619T122818_cis_prep.nc
20181118T120459_cis_prep.nc
20191008T124919_cis_prep.nc

B. ExtremeEarth projection information

PROJCS["Stereographic_North_Pole",
GEOGCS["WGS 84",
DATUM["WGS_1984",
SPHEROID["WGS 84",6378137,298.257223563,
AUTHORITY["EPSG","7030"]],
AUTHORITY["EPSG","6326"]],
PRIMEM["Greenwich",0],
UNIT["degree",0.0174532925199433,
AUTHORITY["EPSG","9122"]],
AUTHORITY["EPSG","4326"]],
PROJECTION["Polar_Stereographic"],
PARAMETER["latitude_of_origin",90],
PARAMETER["central_meridian",0],
PARAMETER["false_easting",0],
PARAMETER["false_northing",0],
UNIT["metre",1],
AXIS["Easting",SOUTH],
AXIS["Northing",SOUTH]]

References

[1] J. Ho, "The implications of Arctic sea ice decline on shipping," *Marine Policy*, vol. 34, no. 3, pp. 713–715, May 2010.

[2] N. Melia, K. Haines, and E. Hawkins, "Sea ice decline and 21st century trans-Arctic shipping routes," *Geophysical Research Letters*, vol. 43, no. 18, pp. 9720–9728, 2016.

[3] L. Pizzolato, S. E. L. Howell, J. Dawson, F. Laliberté, and L. Copland, "The influence of declining sea ice on shipping activity in the Canadian Arctic," *Geophysical Research Letters*, vol. 43, no. 23, pp. 12,146–12,154, 2016.

[4] J. D. Ford, T. Pearce, I. V. Canosa, and S. Harper, "The rapidly changing Arctic and its societal implications," *WIREs Climate Change*, vol. 12, no. 6, p. e735, Nov. 2021, publisher: John Wiley & Sons, Ltd. [Online]. Available: https://doi.org/10.1002/wcc.735

[5] J. Dawson, E. J. Stewart, M. E. Johnston, and C. J. Lemieux, "Identifying and evaluating adaptation strategies for cruise tourism in Arctic Canada," *Journal of Sustainable Tourism*, vol. 24, no. 10, pp. 1425–1441, Oct. 2016, publisher: Routledge. [Online]. Available: https://doi.org/10.1080/09669582.2015.1125358

[6] E. T. o. S. JCOMM, "Sea-Ice Information Services in the World. Edition 2017." World Meteorological Organization, Report, 2017. [Online]. Available: http://dx.doi.org/10.25607/OBP-1325

[7] B. Holt, R. Kwok, and E. Rignot, "Ice Classification Algorithm Development and Verification for the Alaska Sar Facility Using Aircraft Imagery," in *12th Canadian Symposium on Remote Sensing Geoscience and Remote Sensing Symposium*, vol. 2, 1989, pp. 751–754.

[8] F. Hu, G.-S. Xia, J. Hu, and L. Zhang, "Transferring Deep Convolutional Neural Networks for the Scene Classification of High-Resolution Remote Sensing Imagery," *Remote Sensing*, vol. 7, no. 11, pp. 14 680–14 707, Nov. 2015.

[9] R. Pires de Lima and K. Marfurt, "Convolutional Neural Network for Remote-Sensing Scene Classification: Transfer Learning Analysis," *Remote Sensing*, vol. 12, no. 1, p. 86, Dec. 2019, publisher: MDPI AG. [Online]. Available: https://doi.org/10.3390/rs12010086

[10] G. Cheng, X. Xie, J. Han, L. Guo, and G.-S. Xia, "Remote Sensing Image Scene Classification Meets Deep Learning: Challenges, Methods, Benchmarks, and Opportunities," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 3735–3756, 2020, conference Name: IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.

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

[12] N. Hughes and F. Amdal, "ExtremeEarth Polar Use Case Training Data," 2021, publisher: Zenodo Version Number: 1.0.0. [Online]. Available: https://doi.org/10.5281/zenodo.4683174

[13] J. Buus-Hinkler, T. Wulf, A. R. Stokholm, A. Korosov, R. Saldo, L. T. Pedersen, D. Arthurs, R. Solberg, N. Longépé, and M. B. Kreiner, "AI4Arctic Sea Ice Challenge Dataset," Nov. 2022. [Online]. Available: https://data.dtu.dk/collections/AI4Arctic_Sea_Ice_Challenge_Dataset/6244065/2

[14] W. Dierking, "Sea Ice Monitoring by Synthetic Aperture Radar," *Oceanography*, vol. 26, no. 2, pp. 100–111, 2013, publisher: Oceanography Society. [Online]. Available: http://www.jstor.org/stable/24862040

[15] J. Schulz-Stellenfleth and S. Lehner, "Spaceborne synthetic aperture radar observations of ocean waves traveling into sea ice," *Journal of Geophysical Research: Oceans*, vol. 107, no. C8, pp. 20–1–20–19, 2002.

[16] L. Wang, K. A. Scott, L. Xu, and D. A. Clausi, "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*, vol. 54, no. 8, pp. 4524–4533, 2016.

[17] L. Wang, K. A. Scott, and D. A. Clausi, "Sea Ice Concentration Estimation during Freeze-Up from SAR Imagery Using a Convolutional Neural Network," *Remote Sensing*, vol. 9, no. 5, 2017. [Online]. Available: https://www.mdpi.com/2072-4292/9/5/408

[18] C. L. V. Cooke and K. A. Scott, "Estimating Sea Ice Concentration From SAR: Training Convolutional Neural Networks With Passive Microwave Data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 7, pp. 4735–4747, 2019.

[19] A. Stokholm, T. Wulf, A. Kucik, R. Saldo, J. Buus-Hinkler, and S. M. Hvidegaard, "AI4SeaIce: Toward Solving Ambiguous SAR Textures in Convolutional Neural Networks for Automatic Sea Ice Concentration Charting," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[20] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241.

[21] A. Kucik and A. Stokholm, "AI4SeaIce: selecting loss functions for automated SAR sea ice concentration charting," *Scientific Reports*, vol. 13, no. 1, p. 5962, Apr. 2023. [Online]. Available: https://doi.org/10.1038/s41598-023-32467-x

[22] S. Khaleghian, H. Ullah, T. Kræmer, N. Hughes, T. Eltoft, and A. Marinoni, "Sea Ice Classification of SAR Imagery Based on Convolution Neural Networks," *Remote Sensing*, vol. 13, no. 9, p. 1734, Jan. 2021, number: 9 Publisher: Multidisciplinary Digital Publishing

Institute. [Online]. Available: https://www.mdpi.com/2072-4292/13/9/1734

[23] W. Song, H. Li, Q. He, G. Gao, and A. Liotta, "E-MPSPNet: Ice-Water SAR Scene Segmentation Based on Multi-Scale Semantic Features and Edge Supervision," *Remote Sensing*, vol. 14, no. 22, 2022. [Online]. Available: https://www.mdpi.com/2072-4292/14/22/5753

[24] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid Scene Parsing Network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6230–6239.

[25] C. Zhang, X. Chen, and S. Ji, "Semantic image segmentation for sea ice parameters recognition using deep convolutional neural networks," *International Journal of Applied Earth Observation and Geoinformation*, vol. 112, p. 102885, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1569843222000875

[26] H. Boulze, A. Korosov, and J. Brajard, "Classification of Sea Ice Types in Sentinel-1 SAR Data Using Convolutional Neural Networks," *Remote Sensing*, vol. 12, no. 13, p. 2165, Jan. 2020, number: 13 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2072-4292/12/13/2165

[27] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential Deep Learning to Quantify Classification Uncertainty," Montreal, Quebec, Canada, 2018.

[28] N. Asadi, K. A. Scott, A. S. Komarov, M. Buehner, and D. A. Clausi, "Evaluation of a neural network with uncertainty for detection of ice and water in sar imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 1, pp. 247–259, 2021.

[29] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in *34th International Conference on Machine Learning*, Sydney, Australia, 2017, pp. 1321–1330.

[30] X. Chen, K. A. Scott, L. Xu, M. Jiang, Y. Fang, and D. A. Clausi, "Uncertainty-Incorporated Ice and Open Water Detection on Dual-Polarized SAR Sea Ice Imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–13, 2023, conference Name: IEEE Transactions on Geoscience and Remote Sensing.

[31] Q. Yu and D. A. Clausi, "IRGS: Image Segmentation Using Edge Penalties and Region Growing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 12, pp. 2126–2139, Dec. 2008, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[32] P. Yu, A. K. Qin, and D. A. Clausi, "Unsupervised Polarimetric SAR Image Segmentation and Classification Using Region Growing With Edge Penalty," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 4, pp. 1302–1317, Apr. 2012, conference Name: IEEE Transactions on Geoscience and Remote Sensing.

[33] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ser. ICML'16. New York, NY, USA: JMLR.org, 2016, pp. 1050–1059.

[34] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014, publisher: JMLR.org. [Online]. Available: http://dl.acm.org/citation.cfm?id=2627435.2670313

[35] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17, Long Beach, California, USA, 2017, pp. 6405–6416.

[36] C. Dechesne, P. Lassalle, and S. Lefèvre, "Bayesian U-Net: Estimating Uncertainty in Semantic Segmentation of Earth Observation Images," *Remote Sensing*, vol. 13, no. 19, p. 3836, Jan. 2021, number: 19 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2072-4292/13/19/3836

[37] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *2nd International Conference on Learning Representations, ICLR 2014, , April 14-16*, Y. Bengio and Y. LeCun, Eds., 2014. [Online]. Available: http://arxiv.org/abs/1312.6199

[38] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," in *3rd International Conference on Learning Representations, ICLR 2015, May 7-9,*, Y. Bengio and Y. LeCun, Eds., San Diego, CA, USA,, 2015. [Online]. Available: http://arxiv.org/abs/1412.6572

[39] A. Mehrtash, W. M. Wells, C. M. Tempany, P. Abolmaesumi, and T. Kapur, "Confidence Calibration and Predictive Uncertainty Estimation for Deep Medical Image Segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 12, pp. 3868–3878, 2020.

[40] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*, Stanford, CA, USA, 2016, pp. 565–571.

[41] A. Korosov, D. Demchev, N. Miranda, N. Franceschi, and J.-W. Park, "Thermal Denoising of Cross-Polarized Sentinel-1 Data in Interferometric and Extra Wide Swath Modes," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[42] GDAL, *GDAL/OGR Geospatial Data Abstraction software Library*. Open Source Geospatial Foundation, 2022. [Online]. Available: https://gdal.org

[43] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. Alché-Buc, E. Fox, and R. Garnett, Eds., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[44] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous Convolution for Semantic Image Segmentation," *arXiv e-prints*, p. arXiv:1706.05587, Jun. 2017, _eprint: 1706.05587.

[45] K. L. Hermann, T. Chen, and S. Kornblith, "The Origins and Prevalence of Texture Bias in Convolutional Neural Networks," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS'20, Vancouver, BC, Canada, 2020.

[46] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *7th International Conference on Learning Representations, ICLR 2019,*, New Orleans, LA, USA,, 2019.

[47] e. a. Falcon, WA, "PyTorch Lightning," 2019.

[48] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[49] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *ICLR*. San Diego, CA, USA: ICLR, Dec. 2014, p. arXiv:1412.6980.

Rafael Pires de Lima has a bachelor in Geophysics from the University of São Paulo (2011), a MSc. in Geophysics (2017), a second MSc. in Data Science and Analytics (2019), and a Ph.D. in Geophysics (2019) from the University of Oklahoma. His research interests include developing machine learning algorithms for different geospatial and geoscience applications. He also has experience with advanced image processing techniques for seismic attributes development, as well as data analysis for several geophysical methods. He was a Postdoctoral Scientist at the University of Colorado Boulder while writing this manuscript.

Morteza Karimzadeh Morteza Karimzadeh is a PhD in Geography from Penn State (2018), and currently an assistant professor of Geography at the university of Colorado Boulder. He is a spatial data scientist, with use-inspired research in domains such as sea ice mapping, spatial epidemiology, public health, and precision agriculture. His research is human-centered, from algorithmic development to visual analytics and stakeholder engagement.