Accounting for Label Errors When Training a Convolutional Neural Network to Estimate Sea Ice Concentration Using Operational Ice Charts

Manveer Singh Tamber, K. Andrea Scott, *Member, IEEE*, and Leif Toudal Pedersen, *Member, IEEE*

***Abstract*—Convolutional neural networks (CNNs) are being increasingly investigated as a means to extract sea ice concentration from synthetic aperture radar (SAR) in an automated manner. This is often done using ice charts as training data. However, in these charts, an ice concentration label is given to a large region, which may not have a spatially uniform sea ice concentration distribution at the prediction scale of the CNN. This leads to representativity errors, which can be more pronounced at intermediate sea ice concentrations. In this study, we first investigate ways to perturb the ice chart labels to obtain improved predictions to account for the label uncertainty for intermediate ice concentrations. We then propose a method to augment the ice chart data by rescaling the information in the SAR imagery. The method is found to lead to improved accuracy in comparison to using the ice chart labels alone, with accuracy improving from 0.919 to 0.966. The sea ice concentration maps with the augmented labels also have much finer detail than the other approaches evaluated. These details are visually in agreement with expected sea ice concentration from the SAR data.**

***Index Terms*—Convolutional neural network (CNN), ice concentration, synthetic aperture radar (SAR) data.**

I. Introduction

Sea ice concentration is defined as the fraction of a given portion of the ocean surface that is covered by sea ice. It is considered as an essential climate variable by the World Meteorological Organization due to the role it plays in climate and in moderating the heat and momentum transfer at the ocean–ice and ocean–atmosphere interfaces. It is also a key variable in operational ice monitoring, as it is an impediment for ship traffic at high latitudes.

The main source of remote sensing data used for operational ice monitoring is synthetic aperture radar (SAR) imagery. These data are acquired at low frequencies of the microwave spectrum and are, therefore, insensitive to atmospheric moisture. Due to the complexity of the interaction of the SAR signal and the sea ice or ice/snow/ocean system, and the imaging geometry, there is not a straightforward mapping between the signal received by the sensor (backscatter) and the surface properties.

Automated algorithms to estimate sea ice concentration from these images are still under development. At present, SAR images are typically analyzed manually by trained ice analysts employed at national ice services. The products of these analyses, called "ice charts," contain labeled regions, called polygons, that are considered to have spatially homogeneous ice cover. The labels contain the overall concentration of each polygon, as well as the proportion of area covered by up to three main ice types and information of floe size distribution. For practical reasons, polygons are often large compared to the spatial resolution of the SAR data. Common errors in ice charts include operator biases, representativity errors, and uncertainty in setting ice concentration labels for intermediate ice concentrations [1], [2].

Due in part to the manual labor required to generate ice charts, increasing demand for these analyses, and increasing data volumes, there is interest in investigating ways to extract information on the ice cover from SAR data in an automated manner. To this end, several studies have proposed either feature engineering or feature learning approaches [2]–[4]. Here, we focus on feature learning, as it was found to be more suitable in a direct comparison study [5]. Specifically, we focus on learning ice concentration information, although one can also learn information regarding the stage of development of the ice (often referred to as ice type) [6]–[8].

To train convolutional neural networks (CNNs) to predict sea ice concentration from SAR imagery, several previous studies have used ice charts to provide the ice concentration labels [5], [9]. However, as pointed out above, the ice chart labels have errors. In this study, we are mainly concerned with representativity errors. For example, the role of the ice analyst is not to provide polygons that capture all the spatial details of the ice cover. Regions of high ice concentration with some leads (narrow openings) could be assigned a label of 90% ice concentration. At a smaller scale, this label would not be accurate because the pixels in the polygon that are open water are given the same label as pixels in the polygon that are ice.

To address these issues, one approach is to train the network on only "pure" ice and water polygons (those that have only one ice type or are 100% open water [10], and let the CNN infer the information on intermediate ice concentrations. However, given the limited number of pure polygons, to train a CNN effectively

Manuscript received June 3, 2021; revised August 25, 2021; accepted January 3, 2022. Date of publication January 7, 2022; date of current version February 7, 2022. This work was supported in part by ArcticNet—A Network Center of Excellence under Project 62, in part by Environment and Climate Change Canada under Grant GCXE20M002, and in part by the University of Waterloo, President's Research Award. *(Corresponding author: K. Andrea Scott.)*

Manveer Singh Tamber is with the David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: mtamber@uwaterloo.ca).

K. Andrea Scott is with the Department of Systems Design Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: ka3scott@uwaterloo.ca).

Leif Toudal Pedersen is with the DTU Space, Technical University of Denmark, 2800 Kongens Lyngby, Denmark (e-mail: ltp@space.dtu.dk).

Digital Object Identifier 10.1109/JSTARS.2022.3141063

in this manner, one would require a very large database of images. In addition, pure ice polygons could have representativity errors. Giving an automated method, these samples to learn ice concentration without taking this into account could lead to a systematic bias in partial ice concentrations. For example, if the "pure" ice polygons typically have open water within them, the predicted ice concentration could be biased high in marginal ice zones. Another approach is to redraw polygons, although this would be labor intensive, and it may be difficult to achieve polygons that fully represent the details in the ice cover.

In contrast to using only pure ice and water or redrawing polygons, one can use all of the ice chart polygons in training. In this approach, the CNN is tasked with downscaling the coarse-grained ice chart polygons to provide smaller scale ice concentration information. To reduce the impact of representativity errors, one can use the ice concentration directly in a mean absolute error (MAE) loss function, which is less sensitive to outliers than a mean-squared error (MSE) loss function [11]. Alternatively, one can threshold the ice concentration from the ice charts to zeros and ones and train a CNN to predict a probability of ice using a binary cross-entropy (BCE) loss function. Recently, it has been shown that using ice concentration values directly in the BCE loss function, instead of first thresholding, yields improved model predictions [9]. This was attributed to the fact that in such an approach, the label is interpreted as a "soft probability."

In this article, we first build directly on earlier work [9] in which passive microwave (PM) data are used as input to the CNN with the SAR data and explore various ways to accommodate the representativity error associated with the ice chart labels. While the PM data have coarser spatial resolution than SAR (e.g., 5 km–55 km versus 100 m), PM radiometers acquire data over a range of frequencies. Given that the response of the PM signal to various conditions (thin ice, wind roughening, and atmospheric moisture) is frequency dependent [12]–[14], this can enable the PM data to assist the CNN in differentiating between variability in SAR backscatter and/or PM brightness temperatures due to these conditions and variability due to ice concentration. We then propose a novel approach to augment the ice chart labels that is able to account for representativity error and compare this approach to one [2] proposed in an earlier study. We show that using this label augmentation, a significant amount of detail visible in the SAR data can be retained in the ice concentration predictions. The method shows strong generalization capability and is robust to wind roughening of the ocean and other features in the SAR data that are not related to ice concentration.

The rest of this article is organized as follows. We present the database briefly in Section II (an earlier version is discussed in [9]). The methodology is presented in Section III. Section IV presents the experimental setup. Results are presented in Section V. Finally, Section VI concludes this article.

II. Data

The dataset used for this study is the ASIP sea ice dataset—Version 2 [15], produced by the Danish Meteorological Institute (DMI), the Technical University of Denmark, and the Nansen Environmental Remote Sensing Center (NERSC). A corrigendum of this dataset is used, in which nine erroneous scenes were removed. The dataset consists of 452 Sentinel-1 SAR scenes acquired between March 14, 2018 and May 25, 2019 with coregistered corresponding data from Advanced Microwave Scanning Radiometer 2 (AMSR2) and ice charts from the Greenland Ice Service of the DMI. The SAR scenes cover most of the waters surrounding Greenland. For each SAR image, a mosaic of the AMSR2 swath data that gives priority to the swaths acquired closest in time (but before) the SAR acquisition is used. A time window of up to 7 h is considered in selecting the AMSR2 swaths that overlap the SAR imager. The reason why the AMSR2 data are chosen to be before the SAR image acquisition is to "simulate" the operational environment at an ice service. The Sentinel-1 SAR images are C-band dual polarized with a resolution of 93×87 m (range × azimuth) and a pixel spacing of 40×40 m. The AMSR2 data consist of brightness temperatures at seven different frequencies, each with a horizontal and vertical polarization. The instrument field of view for the AMSR2 ranges from 35×62 km for the lower frequency channels to 3×5 km for the higher frequency channels. The AMSR-2 data in the dataset are resampled onto the coordinates of every 50×50 SAR pixels (2×2 km).

Sentinel-1 SAR scenes are corrupted by severe banding or scalloping noise in both the range and azimuth directions. These noise patterns can lead to significant artifacts when geophysical information is derived from the imagery, and it is, therefore, desirable to correct this noise before using the images in automated algorithms. In this study, we use the NERSC noise-corrected data [16], [17] included in the dataset. Additionally, both the SAR and AMSR2 data are normalized before input to the CNN. We normalize the data; as suggested in the ASIP manual [18], the SAR values given as σ^o in decibel were mapped from the approximate range $[-30: +10]$ to the approximate range $[-1: +1]$ by adding 10 and then dividing by 20. To normalize the AMSR2 data, for each brightness temperature value, the mean ($\approx$173.58 K) across all channels was subtracted and then divided by the standard deviation ($\approx$52.68 K) across all channels.

For training, a sliding-window approach was used to extract patches from the SAR scenes. A single patch consists of a $300 \times 300 \times 2$ SAR data array, which is a dual-channel (both HH and HV) 300×300 (12 km × 12 km) image, along with a $6 \times 6 \times 14$ since each AMSR2 pixel is resampled to correspond to 50 SAR pixels. The AMSR2 data consisted of 14 different channels comprising seven frequencies with two polarizations each. For each patch, a corresponding 300×300 sea ice chart label array is extracted from the dataset. In training, the AMSR2 data and the SAR data were the inputs to the model, and the desired model output is the prediction for the sea ice chart label. In the training set, patches were not overlapped. However, in the test set, patches were overlapped by 50 pixels on each side to avoid evaluating the models on predictions produced using padded values, as prescribed in [18]. A 50-pixel edge on each side was discarded from the 300×300 predictions during evaluation.

III. Methodology

A. Training on Ice Chart Labels Directly

Learning sea ice concentration or ice/water information from SAR imagery can be viewed as either a regression over the sea ice concentration values given in the sea ice charts or an ice/water

TABLE I
LABEL PERTURBATION METHODS

Method	Probability of label being set to one
Binary Perturb A	c for $c \in [0, 1.0]$
Binary Perturb B	$2 \times c \times (1 - c)$ for $c \in [0, 0.5]$ or $1 - 2 \times c \times (1 - c)$ for $c \in (0.5, 1]$
Binary Perturb C	$2 \times c^2$ for $c \in [0, 0.5]$ or $1 - 2 \times (1 - c)^2$ for $c \in (0.5, 1]$

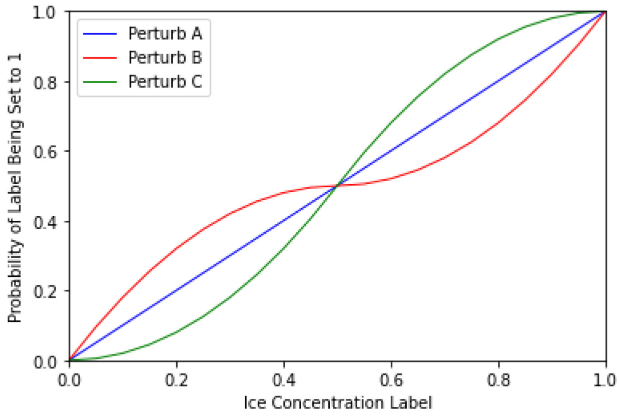


Fig. 1. Probability of the label being set to one for methods, Perturbs A, B, and C, for a given ice concentration label.

classification task. For the latter approach, when a CNN is used, normally the last layer before the loss function is a sigmoid function. This restricts the output to be between 0 and 1, which can be interpreted as an uncalibrated probability, p. This p value is then used in a BCE loss function, where the "true states" are 0s and 1s, obtained by applying a threshold to the ice chart data. In an earlier study [9], improved results were obtained when the sea ice concentration values were used directly as soft probabilities in the BCE loss function, instead of thresholding the ice chart data. We consider this approach as an interpretation of the ice chart label as a probability of the pixel being ice. On this basis, we investigate three approaches to express uncertainty in the ice chart labels.

In method Binary Perturb A, for each pixel in the sea ice chart, the label from the ice chart (a number between 0 and 1) is used as the probability p in a single Bernoulli trial. When training the CNN, the label at that pixel is replaced with either a 0 or 1, where the 0 or 1 is drawn from a Bernoulli distribution with probability p. With regard to training the CNN, these random trials are repeated every single epoch.

Alternatively, the probability of a label being set to one does not have to map linearly to p. Two additional ways of obtaining binary labels from sea ice concentration values are explored. In Binary Perturb B and Binary Perturb C, the probability, p, is related to the ice concentration labels in a different manner (see Table I and Fig. 1). For Binary Perturb B, the probability of setting the label to 1 is higher for lower ice concentrations and lower for high ice concentrations, as compared to Binary Perturb A, whereas the opposite relationship is applied for Binary Perturb C.

We compare the label perturbation methods to using the ice concentration labels more directly in either an MSE or MAE loss function. Due to the representativity error, it is expected that the CNN will interpret many samples as mislabeled, which would correspond to outliers in the loss function. For example, a consolidated ice region with a lead or two (opening in the ice cover) could be given a label of 0.9. If the CNN interprets the open water pixels correctly, these would not match the label of 0.9. These pixels would contribute significantly to the loss if the CNN predicts a zero for these pixels, as it should. Hence, we expect better performance for the MAE loss function as compared to the MSE, as has been found in earlier studies [2], [19].

B. Accounting for Representativity Error I: Mean-Split Loss Function

Mean-split loss was introduced in [2] as a more intuitive way of training a CNN to predict ice concentration using polygon labels from ice charts. Since the polygons labels within patches indicate an average ice concentration over an area, the loss function can be configured to calculate the difference between the mean prediction in the area occupied by a polygon within the patch and the polygon label

$$\mathcal{L}^{\mathrm{MS}}(z, y) = \sum_{i \in I} \frac{M_i}{M} L(i, \bar{y}_i). \tag{1}$$

Given the CNN predictions y and corresponding ice chart concentrations z over a single patch, for each concentration class in the ice concentration labels $I = [0,\ 0.05,\ 0.1,...,1]$, the distance function $\mathcal{L}$ can be used to measure the difference between the concentration label for an area within a patch and the mean prediction over the area occupied by a polygon within the patch. The loss calculated by L is weighted by the number of pixels in the concentration class over the total number of pixels. That is to say, M_i is the number of pixels in a patch with concentration label i, while M is the total number of pixels in the patch. The distance function $\mathcal{L}$ could simply be MAE, where i is the ice concentration category and $\bar{y}_i$ is the mean of the CNN predictions for the given category.

A couple of modifications were made to this function relative to that used in [2]. During training, the loss is computed over a batch of many patches instead of a single patch. This was done so that the true mean concentration over the pixels labeled a certain concentration would be closer to the label given. Since polygon labels cover a large area and ice cover within a given polygon is not completely homogeneous, the true mean ice concentration for a patch within a polygon may differ quite a bit from the ice concentration label given to the polygon. By considering mean-split loss over many patches instead of a single patch, there will be more pixels per concentration class, and thus, the ice concentration label better approximates the true mean ice concentration over these pixels. In addition, it was observed that the predictions of a model trained using this loss function were often much outside of the range [0, 1]. This was a result of the fact that the loss function is only interested in mean predictions over an area, rather than per-pixel predictions. To reduce the occurrence of out of range values, the loss function was modified to punish predictions outside of the range [0, 1]

$$\mathcal{L}^{MS}(z, y) = \sum_{i \in I} \frac{M_i}{M} L(i, \bar{y}_i) + \frac{\sum_{x \in A} (x - 1) + \sum_{x \in B} (-x)}{\alpha M}. \tag{2}$$

In (2), A is the set of all predictions above 1 and B is the set of all predictions below 0. To modify how aggressively the predictions outside of [0, 1] are punished, α values ranging from 0.5 to 8 were tested. A value of $\alpha = 4$ was found to work well in keeping predictions mostly within the range of [0, 1] while scoring relatively well in the R^2 measure discussed later in Section V.

C. Accounting for Representativity Error II: Label Augmentation Using SAR Data

The Binary Perturb and Mean-Split approaches are not directly intended to overcome the representativity error. Here, we present a method to overcome the error and guide our network toward interpreting the smaller scale features in the SAR images, using the SAR data itself. The key assumption is that in a single patch (300 × 300 pixels or 12 km × 12 km) containing both ice and water, ice should generally be associated with higher levels of backscatter as compared to water. This assumption allows us to exploit the SAR data to augment information present in the ice chart labels provided by the DMI ice analysts. We recognize that this assumption will not be valid in all situations. For example, when new ice is present in wind-roughened water, the ice may appear dark, while the water is bright. Additionally, certain atmospheric disturbances (wind or rain) can give rise to strong patterns in SAR imagery that could be misinterpreted as ice information [20]. However, given the strong generalization capability of a CNN, we anticipate that the network will not fixate on these situations. We recognize that this is a shortcoming of the approach and propose to investigate this further through specific approaches for out-of-distribution samples [21]. At this time, since these situations are most pronounced in HH imagery, we carried out experiments using HH and HV separately to augment the labels, in addition to one where they are used together.

In the proposed Augmented Labels method, for the pixels given an ice concentration label of c by the ice analysts, the mean of the augmented labels across these pixels will be approximately equal to c. Higher ice concentration labels will be given to pixels corresponding to brighter pixels in the SAR data, while lower ice concentration labels will be given to pixels corresponding to darker pixels in the SAR data. In the Augmented Labels method, first, 10 × 10 average pooling is performed across both channels of the SAR data to reduce speckle noise. This is followed by bilinear upsampling to convert the dimensions of each SAR image patch back to 300 × 300 pixels (12 km × 12 km). The two images from the two channels are then normalized to have a mean of 0 and a standard deviation of 1, so they can be added together to get a single image patch s. For each concentration class with the exception of 0, 0.95, and 1 (open water, fully compacted but not landfast ice, and landfast ice, respectively), the mean and standard deviation for the corresponding SAR pixels in s given that label are calculated. Then, the standard deviation is multiplied by some uniformity factor. A larger uniformity factor reduces the extent to which the augmented label differs from the original one. Using the modified standard deviation (after multiplication by the uniformity factor) and the mean, a cumulative distribution function (CDF) of the standard normal distribution is evaluated at the value for each SAR pixel in s given the concentration class label being considered. The CDF values will be between 0 and 1 and are centered at 0.5. The CDF quantifies the probability that a sample pixel drawn from the patch that has the given ice concentration label has a backscatter less than or equal to the SAR pixel value from s. If the SAR pixel value has a 50% chance of being greater than a sample pixel value with the same label, then the label for the pixel should be the label given by the ice analyst. If the SAR pixel value has a greater than 50% chance of being greater than a sample pixel value with the same label, then the label for the pixel should be higher than the label given by the ice analyst. This is done by subtracting 0.5 and adding the polygon label to each of the CDF values. Finally, these values are clipped to be between 0 and 1 and become the augmented labels. The Augmented Labels approach is applied separately for each concentration class. If there are not enough pixels with a given concentration label, arbitrarily, if there are less than 10 pixels with a given concentration label, the labels for those pixels are not augmented. The Augmented Labels approach was not applied to labels of 0, 0.95, and 1 since there should be very little representativity error for these labels, with 0 labels being open water, 0.95 labels being ice that is fully compacted, but not landfast, and labels of 1 being landfast ice. A demonstration of the Augmented Labels approach is shown in Fig. 2.

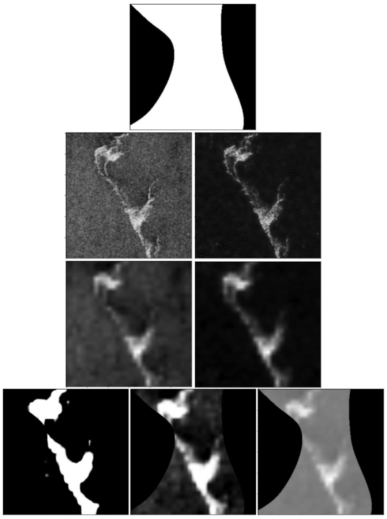


Fig. 2. Demonstration of the SAR augmented labels. Top row: Patch from an interpolated sea ice chart in a 300 × 300 image demonstrating a polygon label, with white representing a label of 0.3 and black representing a label of 0. Second row: HH (left) and HV (right) SAR images corresponding to the patch. Third row: The HH (left) and HV (right) images after the average pooling and bilinear upsampling. Bottom row: Three different examples of augmented labels constructed using uniformity factors of 0.1 (left), 1 (middle), and 10 (right).

For all models trained using Augmented Labels, an MAE loss function was used. Models were initially trained using Augmented Labels testing uniformity factors of 1, 2, and 3. As mentioned above, two additional models were trained using Augmented Labels with a uniformity factor of 1, where the HH and HV channels of the SAR imagery were each used separately to augment the labels. We expect the model with HH augmentation to suffer in conditions with dark ice and light water that can occur during thin ice/strong winds, whereas the model with HV augmentation may suffer from signal to noise issues due to the high noise levels associated with the HV channel.

D. CNN Model

The CNN architecture used was proposed in [9] with the main goal being the fusion of both the AMSR2 brightness temperature data and the SAR data to predict sea ice concentration. Throughout the convolutional layers of the model, same padding was used to retain high-resolution features. By using same padding, through every convolutional layer in the network, the height and width of the image patch are retained at 300 × 300. The CNN begins with six 3 × 3 kernel convolutional layers. After the sixth convolutional layer, four different average pooling layers were used in parallel, each with a window size equal to the receptive field corresponding to the dilation rate of the atrous convolutional layer following the average pooling. This atrous spatial pyramid pooling [22] was used to capture features in the SAR image at different scales. The AMSR2 data pass through bilinear upsampling to match the dimensions of the 300 × 300 SAR data and are concatenated to the output of the four atrous layers along with the output of the second convolutional layer. A final single 1 × 1 convolution is applied to the concatenated data.

The CNN model used was modified from[1] (ASPP_model_extdata_v2 model). Modifications made to this model include increasing the number of filters in the first and second convolutional layers from 12 to 16, increasing the number of filters in the third and fourth convolutional layers from 18 to 20, increasing the number of filters in the fifth and sixth convolutional layers from 18 to 24, and the addition of another convolutional layer right before the final 1 × 1 convolutional layer with 16 1 × 1 filters. In addition, a batch normalization layer was added after the sixth convolutional layer to speed up training, and the dropout layers were removed as they did not seem to improve performance, perhaps because of the model being lightweight with only roughly 58*k* trainable parameters.

For training, a batch size of 32 was used, along with an Adam optimizer at an initial learning rate of 0.001. At the beginning of each epoch, the order of the patches was shuffled. During training, the learning rate was reduced by a factor of 5 if the training loss did not decrease from the minimum by at least 0.001 for five consecutive epochs. Convergence was assumed to have been reached if the learning rate was reduced at least once. Fifty epochs were enough to guarantee that all of the model configurations would converge. At this point, validation loss has plateaued.

IV. Experimental Method

A. Test Data

To test the CNNs, 82 of the 452 scenes (≈18%) were randomly chosen and held out for the test set. Of the remaining 370 scenes, during training, 10% are randomly chosen to form the cross-validation set. From the 82 held out scenes, 67 911 patches were extracted. The ice charts in the 82 scenes consisted of 1182 polygons. For the 13 different ice concentration labels given in the dataset, the number of polygons and pixels per label in the test set is described in Table II.

TABLE II
Polygons and Pixels Per Label in the Test Set

Concentration Label	Number of Polygons	Number of Pixels
0.0	126	2,069,121,203
0.05	19	2,295,952
0.1	69	22,062,161
0.2	107	47,104,895
0.3	66	39,863,514
0.4	52	40,416,337
0.5	39	25,198,123
0.6	57	28,608,902
0.7	76	35,603,471
0.8	79	59,503,511
0.9	121	74,457,492
0.95	147	250,521,441
1.0	224	21,682,998

B. Training Data

It can be seen in Table II that there is a class imbalance in the test data in that the open water labels (those with labels of 0) comprise the majority of the test data. A similar imbalance exists in the training data, with up to ≈73% of training set being open water. To address this imbalance in training the CNN, 60% of the patches where all of the pixels were given a label of 0 were randomly chosen and removed from the training set. This resulted in a roughly 50–50 balance between 0 labels (water) and the remaining 0.05–1.0 labels (ice). The training set consisted of 72 242 patches.

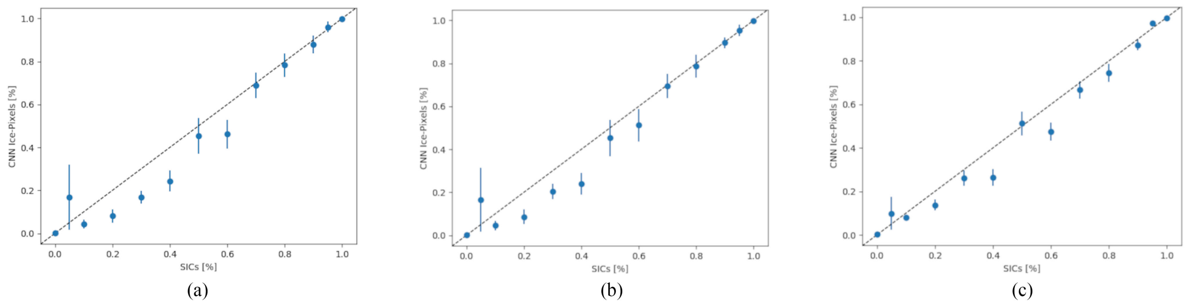
C. Visual Interpretation of CNN Predictions

Visual interpretation is a good tool to look for both gross errors in the predictions that are not captured by metrics and features that the model may capture well. For the former, spurious ice concentration (noise) over what should be open water or model predictions of water over what should be smooth consolidated ice have been noted problems in earlier related studies [23]. For the latter, high-resolution features, such as openings in the ice cover, individual ice floes of a size that can be represented in the SAR imagery, and details of the ice edge, are all features we anticipate the CNN should be able to represent in the model predictions. To visually check the CNN predictions, we investigated the availability of data from Sentinel-2 within a ±1 day time period of the SAR image acquisition. This is a wider time window than is often used, but since we are only doing a visual comparison, we considered this acceptable. Of the four images chosen for visualization, only the image acquired on August 26, 2018 had a good selection of clear sky Sentinel-2 data, as discussed in the following.

V. Evaluation and Results

Herein, for the sake of brevity, the various model configurations will be referred to as follows. There were four models trained on the sea ice chart ice concentration labels directly. The model trained with the BCE loss function will be referred to as

[1] [Online]. Available: https://github.com/damaha/asip-v2/blob/master/keras/models.py



(a)

(b)

(c)

Fig. 3. Percentage of pixels classified as ice for each ice concentration category for (a) BCE, (b) Perturb C, and (c) SARA-1 (both channels). Dots indicate the fraction of ice predicted for the ice concentration bin, and bars indicate the standard error.

TABLE III

R^2 SCORES FOR ICE CONCENTRATION ESTIMATION FOR THE VARIOUS MODELS

Model	R^2 Score	Mean Bias Score	Loss Function
MAE	0.919	-0.026	L_1
BCE	0.925	-0.024	BCE
MSE	0.930	-0.019	L_2
Perturb A	0.929	-0.017	BCE
Perturb B	0.933	-0.010	BCE
Perturb C	0.932	-0.020	BCE
Mean-split	0.954	0.038	Equation (2)
SARA-3	0.952	-0.018	L_1
SARA-1	**0.966**	-0.031	L_1
SARA-1 (HV)	0.965	-0.032	L_1
SARA-1 (HH)	0.963	-0.031	L_1

the BCE model, the model trained with the MAE loss function will be referred to as the MAE model, the model trained with the MSE loss function will be referred to as the MSE model, and the model trained with the mean-split loss function will be referred to as the MS model. For the models trained on binary labels, the model trained on labels produced through the Perturb A method will be referred to as the Perturb A model and likewise for Perturb B and C. Finally, regarding the models trained on the SAR augmented labels, these will be referred to as SARA-#, where # represents the uniformity factor used in the augmentation method. Additionally, the model trained on the labels augmented with the HV SAR channel alone is referred to as SARA-1 (HV), and similarly, the model trained on the labels augmented with the HH SAR channel alone is referred to as SARA-1 (HH). These models were trained using a uniformity factor of 1, which was sufficient to illustrate the impact of using only the HH or HV channel in the SAR augmented label approach.

A. *Quantitative Comparison of Models*

Our first comparison quantitatively compares the model predictions with the ice chart labels for the test dataset. Given the representativity errors, a per-pixel accuracy measure was avoided because the sea ice charts do not contain useful per pixel labels. Instead, an R^2 measure was adopted. The CNN predictions were thresholded at an ice concentration of 0.5. Predictions above the threshold were considered ice and predictions below the threshold were considered water. For each ice concentration bin in the ice charts, we then add up the total number of pixels classified as ice and the total number of pixels classified as water. The fraction of pixels classified as ice over the total number of pixels given a certain ice concentration label should roughly equal the ice concentration value. For each model, we then have a relationship between the fractions estimated and the ice concentration bin, from which an R^2 value can be determined. Additionally, a mean bias score is included. For each concentration label i in $[0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 1]$, the bias for the given label i is first calculated as difference between the sum of all predictions for the given label and the label value. The mean bias score is the sum of these biases divided by the total number of ice concentration labels, which is 13. These scores are given in Table III for each model.

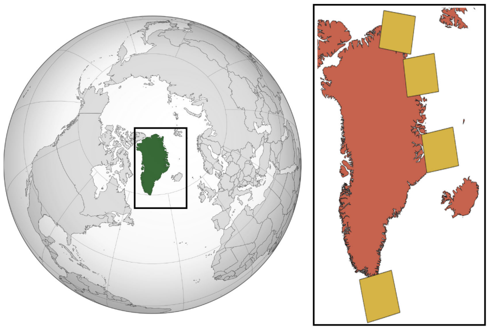


Fig. 4. Location of the four scenes from the test that are examined in detail. These scenes were chosen due to the variety of conditions they represent. From top to bottom are the scenes from August 26, 2018, September 10, 2018, March 22, 2018, and March 14, 2018. Coastline obtained from [Wessel and Smith shoreline database].

It can be seen that the R^2 values are the highest for the SARA-1 models. As the uniformity factor is increased, and the impact of the label augmentation is decreased, the R^2 scores decrease with SARA-3 and have a lower R^2 score than SARA-1. The model trained using the mean-split loss function also performed relatively well as this loss function attempts to overcome the representativity error. The MAE, BCE, MSE and the Perturb models score within the same neighborhood with the Perturb B model scoring the highest and the MAE model scoring the lowest. However, of these six models, the MAE model visually produces the fewest spurious ice predictions over open water.

The mean bias scores presented in Table III indicate that the bias is the lowest for the MSE and Perturb models and slightly greater in magnitude (more negative) for the SARA-1 models. This can be understood from looking at Fig. 3, where it can

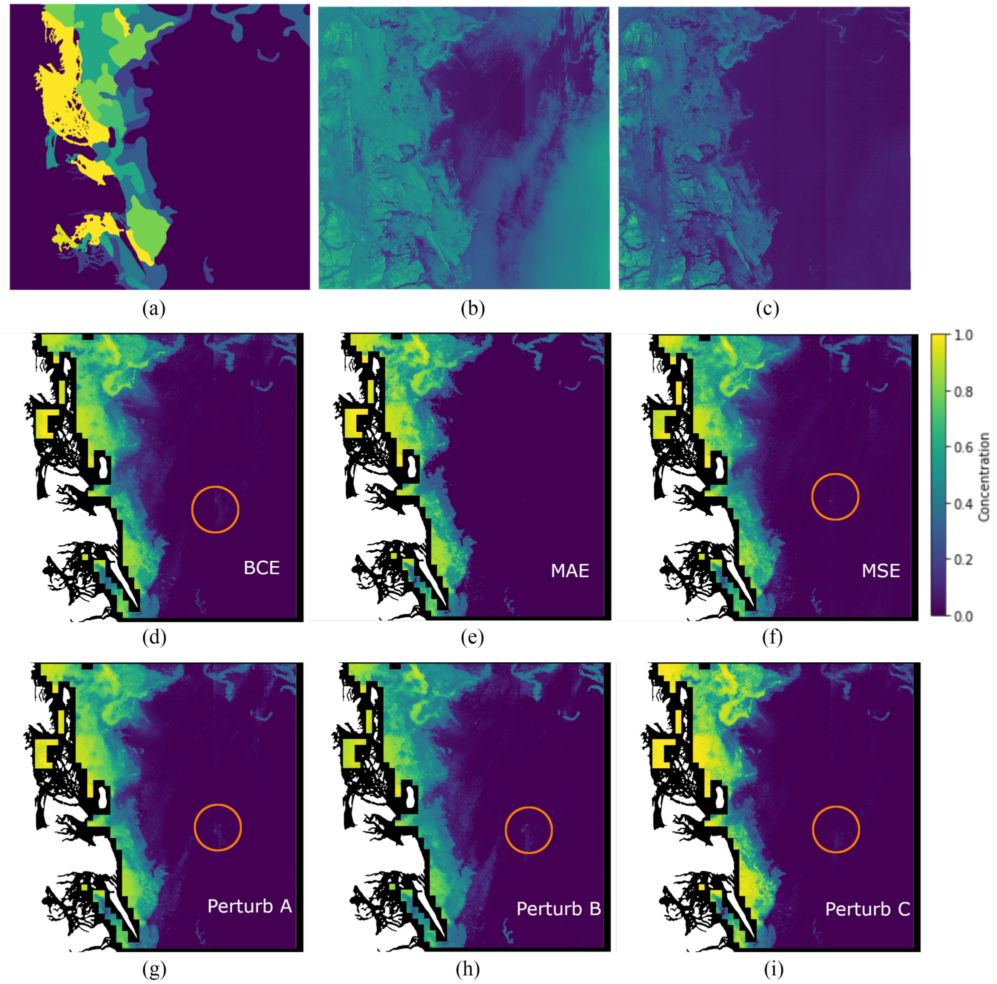


Fig. 5. Scene acquired September 10, 2018, 08:18:14 UTC, covering Greenland's northeast coast. Central latitude and longitude: 77.4° N, 15.4° W. (a) Sea ice chart. (b) HH SAR image. (c) HV SAR image. (d) BCE model predictions. (e) MAE model predictions. (f) MSE model predictions. (g) Perturb A model predictions. (h) Perturb B model predictions. (i) Perturb C model predictions. The orange circle indicates regions where in some cases, spurious ice is retrieved over the open water.

be seen that there is a larger positive bias for the Perturb C model than SARA-1 for the open water category, which would partially compensate for the negative bias associated with the other ice concentration categories. Similarly, the mean bias score is positive for the Mean-Split model, which is in agreement with what we see qualitatively in the results (see Figs. 6 and 7).

To calculate the significance of the R^2 score and the mean bias score, a bootstrapping approach was used, where each sample consisted of a fixed number of scenes (e.g., 82) chosen from the test set with replacement. A number of trials (e.g., 10, 30, and 50) were then carried out, where, for each sample, an R^2 value and a mean bias value were calculated. This was repeated for each model, yielding a set of R^2 estimates. It was found that the results were repeatable, with the set of SARA models consistently outperforming the others, and with the SARA-1 models having consistently ≈ 0.01 higher R^2 scores than the others.

An alternative approach to compare models was also carried out, which allows an estimate of the scatter of model predictions within each ice concentration bin. In this approach, for each scene and for each concentration label i in $[0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 1]$, the number of pixels predicted by the model as ice is added up to give a sample n_i for concentration bin i. Provided that the number of pixels in the given concentration bin in the ice chart m_i is nonzero, n_i/m_i becomes a sample for the proportion of pixels labeled i that are predicted as ice. Then, for any i, the mean of all of the samples n_i/m_i should be equal to i and an R^2 score can be calculated. The standard error is considered for each label i and is calculated as the standard deviation of the samples n_i/m_i divided by the square root of the number of samples, which is equal to the number of scenes that contain the label i.

The performance of the models was similar using this approach, with SARA-1 having the highest R^2 values over the widest range of thresholds. In comparing the scatter in model predictions for the various ice concentration bins, it can be seen that the scatter is greatest for the 0.05 ice concentration bin for all models shown (and all models tested). As compared to the BCE models, incorporating perturbed labels, using Perturb C, can be seen to improve the estimates for intermediate ice concentration categories, although the scatter for the individual concentration bins is similar. When the augmented labels are used, shown in

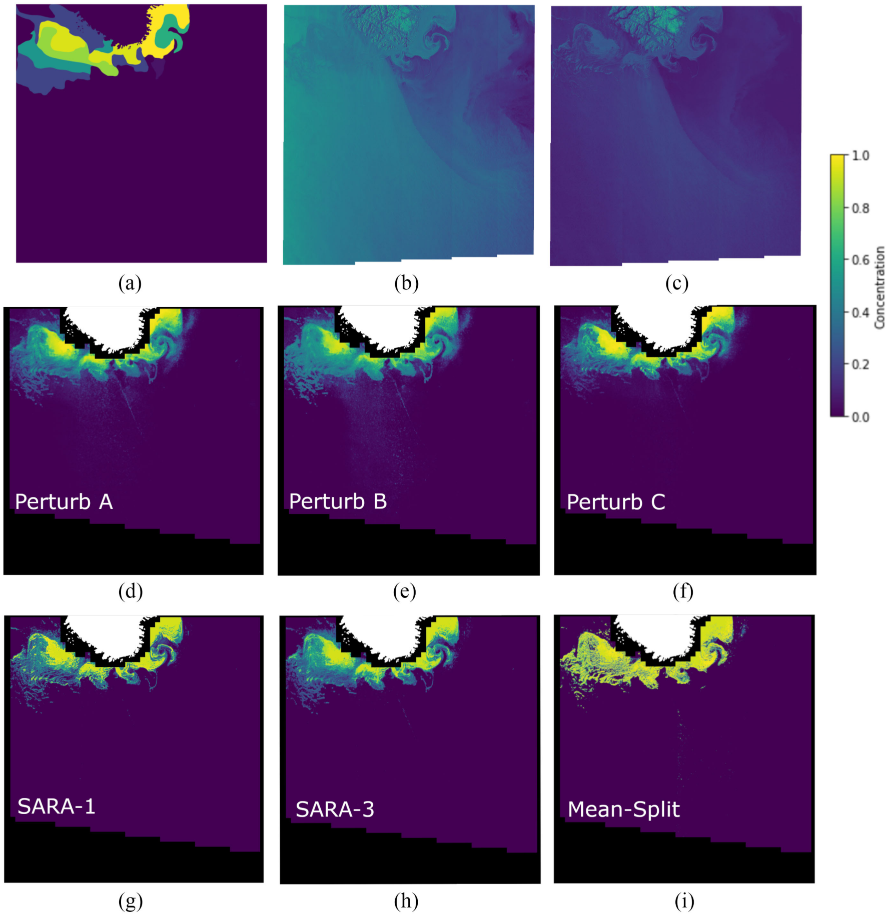


Fig. 6. Scene acquired March 14, 2018, 20:27:22 UTC, covering Cape Farewell, Greenland. Central latitude and longitude: 58.6° N, 42.7° W. (a) Sea ice chart. (b) HH SAR image. (c) HV SAR image. (d) Perturb A model predictions. (e) Perturb B model predictions. (f) Perturb C model predictions. (g) SARA-1 model predictions. (h) SARA-3 model predictions. (i) Mean-Split model predictions. Note that Perturb B appears to slightly underpredict the high ice concentrations, but is able to show detail for low ice concentrations, which is expected based on Fig. 1. Also note that there is significant wind roughening on this date, which is more visible in the HH image than in the HV image, although the HV backscatter is slightly elevated.

Fig. 3(c), the scatter in the individual ice concentration bins is reduced (the vertical bars for intermediate ice concentration bins are smaller), and the bias for the 0.05 concentration bin is reduced, although the bias for the concentration bins 0.6–0.8 is slightly higher.

B. *Visual Comparison of Model Predictions for SAR Scenes*

We now compare methods through visualization of scene-level predictions. For this purpose, four SAR scenes were chosen (locations shown in Fig. 4) that represent a variety of ice and open water conditions. In the visualizations, white is used to represent land, and black is used to represent masked out areas (areas for which predictions are not generated). The predictions are shown as values between 0 and 1. For the models where predictions could have been outside of the range [0, 1], the predictions were clipped to be within this range. In Fig. 5, the Perturb A, B, and C models are compared with the MAE, BCE, and MSE models. Shown in the orange circles drawn over the BCE, MSE, and Perturb A, B, and C predictions is some noise over what should be predicted as open water. This noise seems to result from a combination of wind roughening in the HH SAR image and subswath striping in both SAR images, although it is most pronounced in the HV image. This noise is not present in the MAE model predictions. The light blue patterns over what is actually water signifying predictions of ice tend to occur closer to the ice.

In Fig. 6, the Perturb A, B, and C models are compared as well as SARA-1, SARA-3, and Mean-Split. The predictions look fairly similar for the Perturb models. Every Perturb model prediction seems to show some noise over water closer to the ice edge, with Perturb C being the least noisy. It can also be seen that Perturb B is able to capture the low concentration ice, but underpredicts the magnitude of higher ice concentration, while Perturb C has the opposite tendencies. This is in agreement with the relationship shown in Fig. 1. The SARA-1 and SARA-3 predictions are fairly similar to those from the Perturb models with better resolved spatial features. There is slightly enhanced ice concentration in SARA-1 toward the edge of the marginal ice area in the western portion of the image, possibly due to elevated HV backscatter in this region [see Fig. 6(c)]. The Mean-Split model appears to overpredict the ice concentration in the region on the ice chart indicating concentrations in the range of 0.1–0.7.

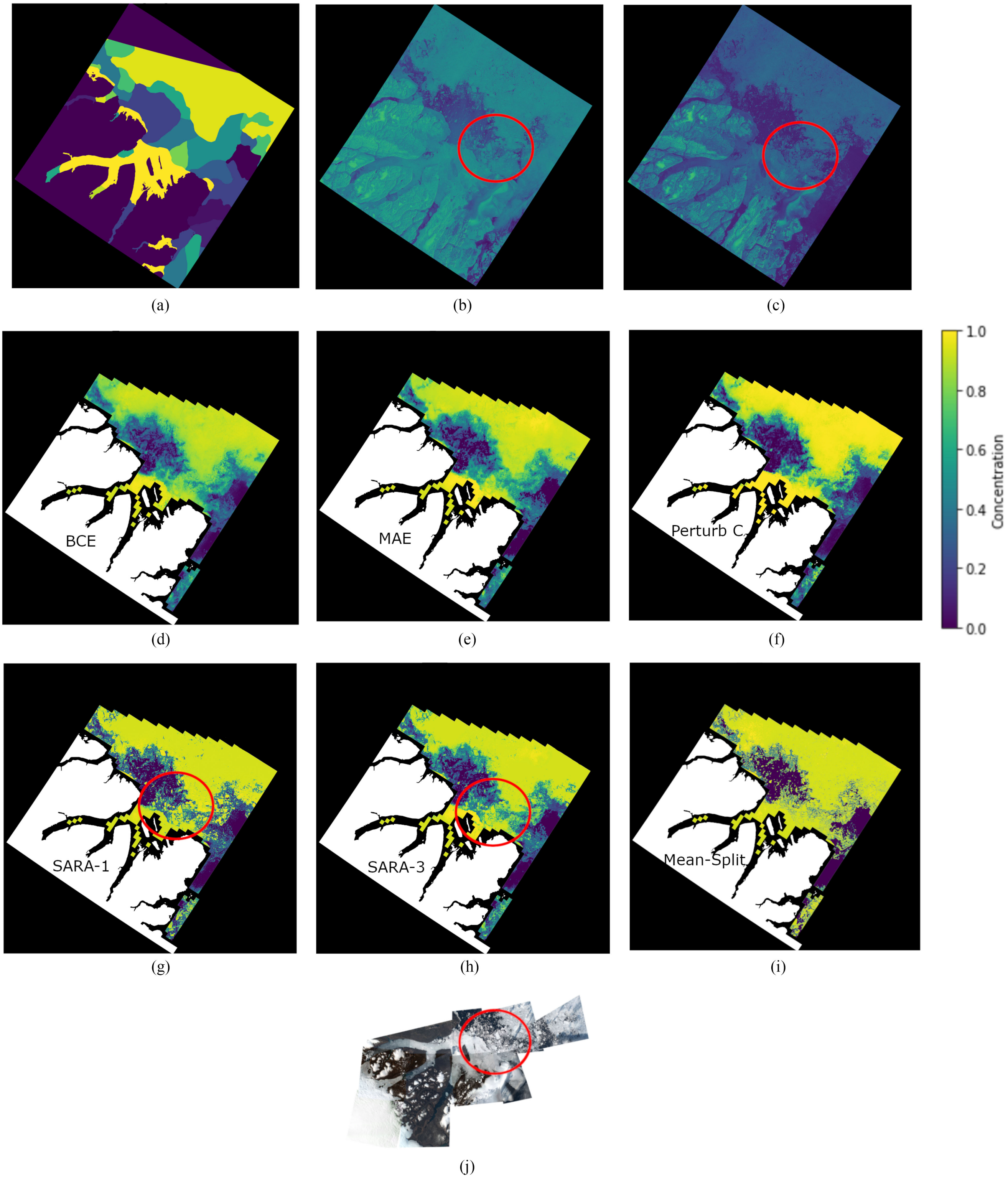


Fig. 7. Scene acquired August 26, 2018, 09:30:44 UTC, covering Greenland's northeast coast. Central latitude and longitude: 82.3° N, 18.7° W. All images are reprojected onto EPSG:32627. (a) Sea ice chart. (b) HH SAR image. (c) HV SAR image. (d) BCE model predictions. (e) MAE model predictions. (f) Perturb C model predictions. (g) SARA-1 model predictions. (h) SARA-3 model predictions. (i) Mean-Split model predictions. (j) Mosaic of Sentinel 2 imagery (RGB: 665, 560, and 490 nm) acquired on August 26, 2018 ±1 day. The SARA and Mean-Split method are able to pick up the small floes that can be seen in the SAR imagery, as indicated by the region shown in the red circle.

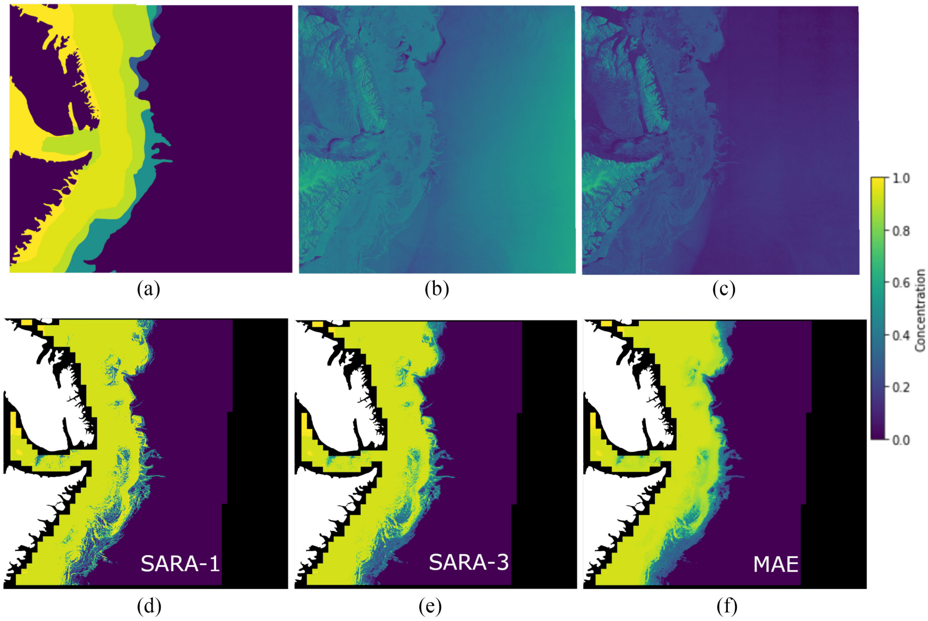


Fig. 8. Scene acquired March 22, 2018, 08:04:25 UTC, covering Greenland's central east coast. Central latitude and longitude: 70.1° N, 19.7° W. (a) Sea ice chart. (b) HH SAR image. (c) HV SAR image. (d) SARA-1 model predictions. (e) SARA-3 model predictions. (f) MAE model predictions.

In Figs. 7 and 8, the SARA models' predictions are shown. Considering Fig. 7, as shown within the red circles, the models trained on the augmented labels can predict relatively smaller individual ice floes, with SARA-1 separating the ice floes most clearly. An RGB mosaic from Sentinel-2 imagery is shown in Fig. 7(j). If one considers some of the larger individual floes in the red circle, SARA-1 appears able to better reconstruct these floes in the manner that they contrast with the underlying distribution of smaller floes. While SARA-3 and Mean-Split do show some floe structure as well, for SARA-3, the results are a little blurry, while for Mean-Split, the ice cover appears overestimated in some places and floes are hard to distinguish. The other models fail to identify individual floes clearly in their predictions. Instead, wherever distinct ice floes are present, the other models predict ice in the surrounding waters as well, extending the predicted ice edge. This is perhaps an artifact of the representativity error in the labels. Considering Fig. 8, the SARA models place a greater emphasis on features near the ice edge that may represent eddy activity.

C. Patch Visuals

To further illustrate the methodology, we show patches extracted from the SAR images along with the augmented labels and model predictions. One concern when devising the SAR label augmentation method was that in this method, the CNN could be extra sensitive to noise from the SAR images, because this noise would be in the label. Nonetheless, considering the patch observed in the first column in Fig. 9, although there is what appears to be wind roughening in the HH SAR image, this noise does not show in the SARA model predictions. Similarly with Fig. 7, on the right edge of the scene, there appears to be wind roughening over an opening in the ice. Again, the SARA models are resilient to this noise and do not reflect this noise in the predictions. The SARA-1 model is also able to capture various small-scale details that can be seen in the SAR images, correctly inferring them as ice or water, respectively.

The fact that the BCE, MAE, MSE, and Perturb A, B, and C models are predicting ice on water surrounding ice is clearer in the patch visuals in Fig. 9 since a patch represents a zoomed-in view. The model trained on the augmented labels produces sharp ice edges in its predictions with the true ice edges being apparent in the SAR images. The other models have a blurry appearance. Even when the predictions from these models are thresholded, these models fail to sharply capture the ice edge while separating the ice from the water. Aside from the SARA models, the MS model is unique in that it does not blur predictions, avoiding nonzero predictions over water. Predictions from the MS model are very binary in nature, and it is, therefore, able to produce good predictions of intermediate ice concentrations at a fine scale.

VI. Conclusion

In this study, we developed a novel method to improve the representation of fine scale details in CNN predictions of sea ice concentration from SAR imagery when ice charts are used to provide the training labels. The method is surprisingly robust given the wide range of SAR signatures, from smooth ice to wind roughened regions, and leads to improved predictions of sea ice concentration, particularly for intermediate ice concentrations. Future work will look at the sensitivity of the method over a wider geographic region that covers a broader range of ice conditions, in addition to more in-depth comparisons with ice concentration from PM sensors and optical data. We will also more rigorously investigate the underlying assumption of higher backscatter values corresponding to ice by investigating the ability of our network to identify samples that do not fit this assumption as out-of-distribution samples. We also plan to

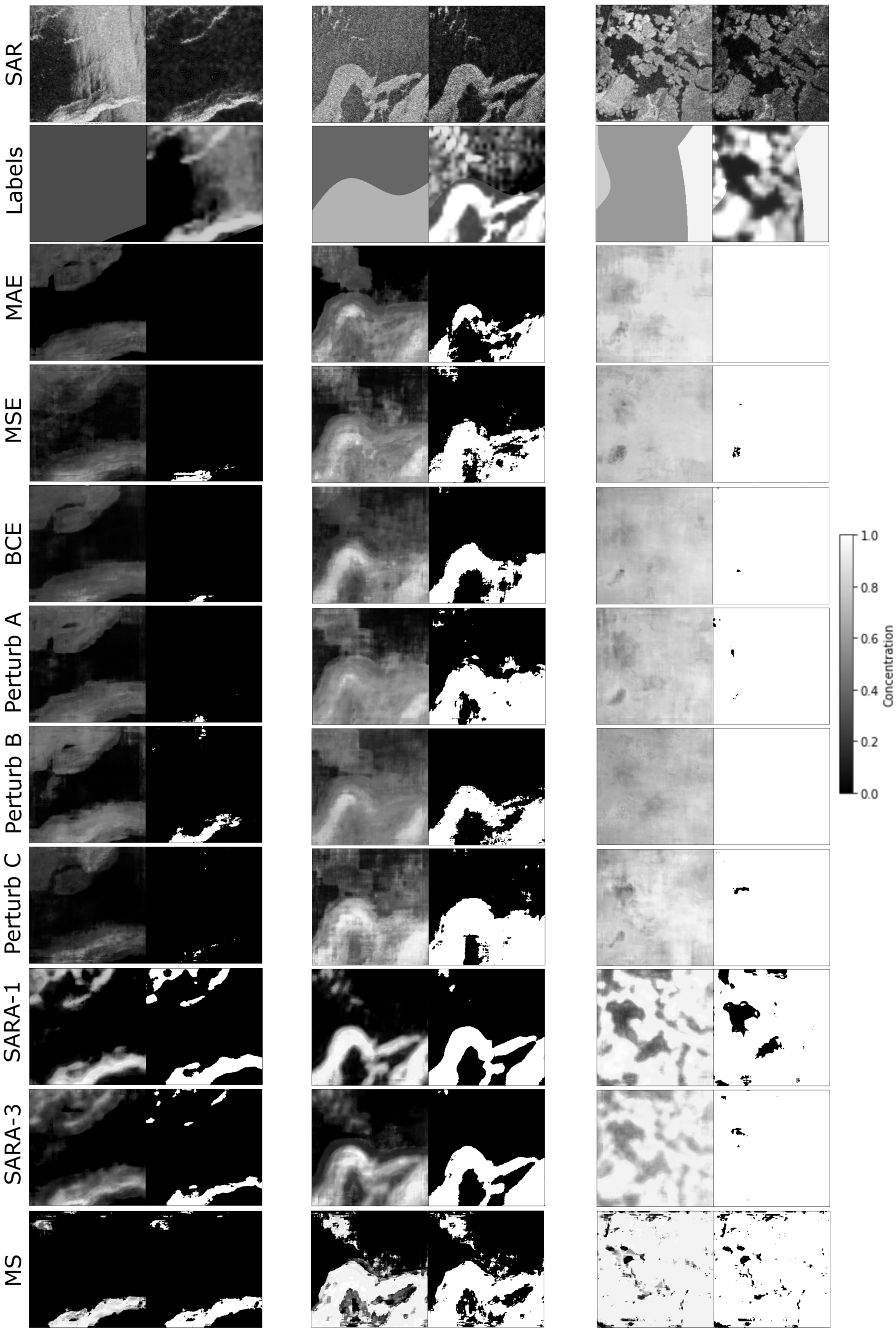


Fig. 9. Three sets of patches. Each set of two columns represents a patch of 300 × 300 pixels corresponding to 12 × 12 km. The first patch (leftmost columns) is from a scene acquired on September 10, 2018, from northeast Greenland, the second patch (middle columns) is from a scene acquired on March 30, 2019, from southwest Greenland, and the third patch (rightmost columns) is from a scene acquired on December 12, 2018, from northwest Greenland. SAR: on the left is the HH SAR image and on the right is the HV SAR image. Labels: on the left are the SIC labels and on the right are the SAR augmented labels (uniformity factor = 1). The remaining nine rows are the predictions for nine model configurations. For each prediction, on the left are the predictions between 0 and 1 and on the right are the predictions after being thresholded at 0.5. Note that the concentration bar on the right does not apply to the SAR images.

address the question of what kind of distribution should be used for the Augmented Labels approach. In this study, our CDF was based on a Gaussian distribution for several reasons. For example, our data are multilooked, and even after this, average pooling is applied. In addition, this study represents a first attempt at the Augmented Labels concept; hence, the additional complexity of a different distribution did not seem warranted. Finally, the CNN did not experience any difficulty converging with this assumption, likely because it is fairly simple and is parameterized by only a mean and a variance. However, we note that the K-distribution is often used as a statistical model for SAR data [24], although perhaps not the same distribution for all ice types [25]. The impact of this choice on CNN convergence and model predictions for a range of ice conditions will be investigated in a future study.

Acknowledgment

The authors would like to thank Compute Canada for providing computing resources.

References

[1] J. Karvonon, J. Vainio, M. Marnela, P. Eriksson, and T. Niskanen, "A comparison between high-resolution EO-based and ice analyst-assigned sea ice concentrations," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 4, pp. 1799–1807, Apr. 2015.

[2] L. Wang, "Learning to estimate sea ice concentration from SAR imagery," Ph.D. dissertation, Dept. Syst. Des. Eng., Univ. Waterloo, Waterloo, ON, Canada, 2016.

[3] J. Karvonen, "Baltic sea ice concentration estimation based on C-band HH-polarized SAR data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 6, pp. 1874–1884, Dec. 2012.

[4] Q. Yu and D. Clausi, "SAR sea-ice image analysis based on iterative region growing using semantics," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 12, pp. 3919–3931, Dec. 2007.

[5] L. Wang, K. Scott, and D. Clausi, "Sea ice concentration estimation during freeze-up from SAR imagery using a convolutional neural network," *Remote Sens.*, vol. 9, no. 5, 2017, Art. no. 408.

[6] R. Kruk, C. Fuller, A. Komarov, and D. Isleifson, "Proof of concept for sea ice stage of development for classification using deep learning," *Remote Sens.*, vol. 12, no. 15, 2020, Art. no. 2486.

[7] H. Boulze, A. Korosov, and J. Brajard, "Classification of sea ice types in Sentinel-1 SAR data using convolutional neural networks," *Remote Sens.*, vol. 12, no. 15, 2020, Art. no. 2165.

[8] W. Song, M. Li, Q. He, D. Huang, C. Perra, and A. Liotta, "A residual convolutional neural network for sea ice classification with Sentinel-1 SAR imagery," in *Proc. IEEE Int. Conf. Data Mining Workshops*, 2018, pp. 795–802.

[9] D. Malmgren-Hansen *et al.*, "A convolutional neural network architecture for Sentinel-1 and AMSR2 data fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 1890–1902, Mar. 2021.

[10] A. Komarov and M. Buehner, "Ice concentration from dual-polarization SAR images using ice and water retrievals at multiple spatial scales," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 950–961, Feb. 2021.

[11] P. Rousseeuw and A. Leroy, *Robust Regression and Outlier Detection*. Hoboken, NJ, USA: Wiley, 1987.

[12] K. Naoki, J. Ukita, F. Nishio, M. Nakayama, J. Comiso, and A. Gasiewski, "Thin sea ice thickness as inferred from passive microwave and in situ observations," *J. Geophys. Res.*, vol. 113, 2008, Art. no. C02S16.

[13] K. Scott, M. Buehner, A. Caya, and T. Carrieres, "Direct assimilation of AMSR-E brightness temperatures for estimating sea ice concentration," *Monthly Weather Rev.*, vol. 140, no. 3, pp. 997–1013, 2012.

[14] P. Hwang, "Foam and roughness effects on passive microwave remote sensing of the ocean," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 8, pp. 2978–2985, Aug. 2012.

[15] R. Saldo, M. B. Kreiner, J. Buus-Hinkler, L. T. Pedersen, D. Malmgren-Hansen, and A. A. Nielsen, *AI4Arctic/ASIP Sea Ice Dataset—Version 2*. Kongens Lyngby, Denmark: Tech. Univ. Denmark, 2020.

[16] J. Park, A. Korosov, M. Babiker, S. Sandven, and J. Won, "Efficient noise removed for Sentinel-1 TOPSAR cross polarization channel," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1555–1565, Mar. 2018.

[17] J. Park, J. Won, A. Korosov, M. Babiker, and N. Miranda, "Textural noise correction for Sentinel-1 TOPSAR cross polarization channel images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 4040–4049, Jun. 2019.

[18] M. Kreiner, J. Hinkler, L. Pedersen, D. Hansen, M. Babiker, and A. Korosov, *ASID-V2, AI4Arctic/ASIP Sea Ice Dataset—Version 2 User Manual*, 1st ed. Copenhagen, Denmark: Danish Meteorol. Inst., 2020.

[19] C. L. Cooke and K. A. Scott, "Estimating sea ice concentration from SAR: Training convolutional neural networks with passive microwave data," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4735–4747, Jul. 2019.

[20] K. Topouzelis and D. Kitsiou, "Detection and classification of mesoscale atmospheric phenomena above sea in SAR imagery," *Remote Sens. Environ.*, vol. 160, pp. 263–272, 2015.

[21] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *Proc. Int. Conf. Learn. Represent.*, 2017.

[22] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, Atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[23] K. Radhakrishanan, A. Scott, and D. Clausi, "Sea ice concentration estimation: Using passive microwave and SAR data with a U-net and curriculum learning," *IEEE J. Sel. Topics Earth Observ. Appl. Remote Sens.*, vol. 14, pp. 5339–5351, 2021.

[24] I. Joughin, D. Winebrenner, and D. Percival, "Probability density functions for multilook polarimetric signatures," *IEEE Trans. Geosci. Remote Sens.*, vol. 32, no. 3, pp. 562–574, May 1994.

[25] M. Collins, C. Livingstone, and R. Raney, "Discrimination of sea ice in the Labrador marginal ice zone from synthetic aperture radar image texture," *Int. J. Remote Sens.*, vol. 18, pp. 535–571, 1987.

Manveer Singh Tamber is working toward the undergraduate degree with the David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, ON, Canada.

His research interests include applied machine learning and artificial intelligence research.

K. Andrea Scott (Member, IEEE) received the B.A.Sc. degree from the University of Waterloo, Waterloo, ON, Canada, in 1999, the M.A.Sc. degree from McMaster University, Hamilton, ON, in 2001, and the Ph.D. degree from the University of Waterloo in 2008, all in mechanical engineering.

She was a Postdoctoral Researcher with the Data Assimilation and Satellite Meteorology Research Section, Environment and Climate Change Canada, Toronto, where she was part of a team involved in the development of a sea ice data assimilation system. In 2012, she joined the Department of Systems Design Engineering, University of Waterloo, as a Faculty Member with a specialization in sea ice remote sensing and data assimilation.

Leif Toudal Pedersen (Member, IEEE) was born in Denmark in 1957. He received the M.S. degree in microwave engineering and the Ph.D. degree in passive microwave remote sensing of sea ice from the Technical University of Denmark (DTU), Kongens Lyngby, Denmark, in 1982 and 1992, respectively.

From 1982 to 2000, he was a Research Assistant with the Electromagnetics Institute, DTU. From 2000 to 2007, he was an Associate Professor with Oersted-DTU, Denmark. From 2007 to 2017, he was a Senior Researcher with the Danish Meteorological Institute, Copenhagen, Denmark. Since early 2017, he has been a part-time Senior Researcher with the DTU Space, DTU. He part-time runs his own company eolab.dk. His research interests include the retrieval of ice, ocean, and atmospheric parameters from multispectral microwave radiometer measurements and other methods for remote sensing of sea ice.

Dr. Pedersen has been a member of the Danish National Committee for Climate Research, the Danish National Committee for the International Polar Year, and the Danish National Committee for Scientific Committee on Antarctic Research (SCAR). He was a Danish delegate to European Space Agency (ESA)'s Program Board for Earth Observations and has served on ESA's Sentinel-1 Mission Advisory Group.