Evaluation of a neural network with uncertainty for detection of ice and water in SAR imagery

Nazanin Asadi, K. Andrea Scott *Member IEEE*, Alexander S. Komarov *Member IEEE*, Mark Buehner and David A. Clausi *Senior Member IEEE*

***Abstract*—Synthetic aperture radar (SAR) sea ice imagery is a promising source of data for sea ice data assimilation. Classification of SAR sea ice imagery into ice and water is of particular relevance due to its relationship with ice concentration, a key variable in sea ice data assimilation systems. With increasing volumes of SAR data, automated methods to carry out these classifications are of particular importance. While several automated approaches have been proposed, none look at the impact of including an estimate of uncertainty of the model parameters and input features on the classification output. The present study uses an established database of SAR image features to train a multilayer perceptron (MLP) neural network to classify pixel locations as either ice, water or unknown. The classification accuracies are benchmarked using a recently developed logistic regression approach for the same database. The two methods are found to be comparable. The MLP approach is then enhanced to allow uncertainty to be estimated at each pixel location. Following methods proposed in the deep learning community, two kinds of uncertainty are considered. The first, epistemic uncertainty, is that due to uncertainty in the MLP weights. The second kind of uncertainty, aleatoric uncertainty, is that which cannot be explained by the model, and is therefore associated with the input data. It is found that including these uncertainties in the MLP models reduces their accuracies slightly, but also reduces misclassification rates. This is of particular importance for data assimilation applications, where misclassifications could severely degrade the analysis.**

***Index Terms*—sea ice, synthetic aperture radar, classification, neural network, uncertainty, data assimilation**

I. Introduction

Synthetic aperture radar (SAR) imagery is widely used by operational ice services to provide information on ice conditions to users, such as shipping companies and local communities. The information provided often comes in the form of ice charts. These are maps where the ice cover in specified regions is given a label indicating the ice concentration of the region and the dominant ice types (e.g., new ice, first-year ice, multi-year ice). Ice charts are generated manually based on visual inspection of SAR imagery, in addition to imagery from passive microwave and visual infrared sensors, ship reports and other in-situ data sources. Constructing these charts is time consuming, and the resulting maps may contain errors due to the subjective nature of their preparation [**?**]. However, they are still considered the most accurate source of information regarding ice conditions [**?**] that is available on a near-daily basis. With the current Sentinel-1 mission, volumes of SAR imagery available for analysis have increased considerably, and will continue to do so for the near future with the new RADARSAT constellation mission. It is desirable to have a process in place that can either fully or partially automate the analysis of this imagery. For example, a process to automate analysis of images that are then straightforward to interpret is desirable so that operators can focus efforts on interpreting images with more challenging ice conditions.

Significant progress has been made in automating analysis of SAR sea ice imagery, with a common application being classification of the imagery into ice and water. For this task a variety of methods have been proposed, such as decision trees [**?**], logistic regression [**?**], neural networks [**?**], support vector machines [**?**] and a method combining a support vector machine with an iterative region growing image segmentation [**? ?**]. An overview of these approaches is given in a recent review [**?**]. Some of these approaches carry out pixel-wise classification, while others perform image segmentation first, before labelling the segments into distinct classes. The accuracies of these methods are typically fairly high, with common problems being classification during the summer, misclassification of new ice as open water, and unknown accuracy of the test data.

None of the methods presented in the literature thus far have investigated the contribution of the model parameters separately from the contribution of the input features to the uncertainty in the model prediction. For the present study we chose a neural network (NN) to carry out ice/water classification of SAR imagery, and we explore the applicability of recent work on different types of uncertainty for NN applications [**? ?**], to this problem domain. We compare our results directly with a recently developed logistic regression classifier that has carefully chosen SAR features as input. Specifically, we seek to answer the following questions

- Can the neural network provide comparable classification accuracies to the logistic regression method?
- How do the accuracies and misclassification rates change when uncertainty estimates are taken into account?
- How are the uncertainty estimates from the NN related to the features used as input to the network?

II. Background

Artificial neural networks (NN) are a powerful machine learning tool. They have been widely used in remote sensing problems, including sea ice applications [**? ? ?**], over the past 20 years. Neural networks are in the category of nonparametric

N. Asadi, K.A. Scott and D.A. Clausi are with the Department of Systems Design Engineering at the University of Waterloo, Waterloo, Canada (email: ka3scott@uwaterloo.ca)

A. Komarov and M. Buehner are with the Data Assimilation and Satellite Meteorology Research Section at Environment and Climate Change Canada, Dorval, Canada

machine learning methods i.e., they do not require any specific assumption about the statistical distribution of the mapping function [**?**]. An artificial neural network is a network constructed by layers of elements called artificial neurons. Each neuron receives a set of inputs and produces an output using a specified (nonlinear) activation function. The network is constructed by feeding the output from one layer of neurons as input to the next layer of neurons. Learning the values of the connections between these layers, called weights, defines the optimization problem to be solved in the training of the neural network. Due to the nonlinearity of the activation functions, and the high-dimensional nature of the problem, finding an optimal solution for a NN (and knowing when it is optimal for a given data set) is a challenging optimization problem [**?**].

There is no unique set of values for the weights that can describe the learning rule, and some uncertainty in the values of the weights is expected. In the NN community this is considered the *epistemic* uncertainty, which is the uncertainty that can be reduced by a better model, or more data. This term has its origins in the risk community [**?**]. To capture this uncertainty, it has been proposed [**?**] to generate a set of NN model outputs with each output varying from the others in terms of which weights are active when the prediction is made (i.e., the model outputs differ in the weights that have been randomly 'dropped' from the network [**?**]). The uncertainty is considered as either the variance of the model output for a regression problem [**?**] or the entropy of the mean probability distribution of the model output for a classification problem [**?**]. This has been coined Monte Carlo dropout, or MC dropout. The relationship between MC dropout and Bayesian NNs has been discussed in recent literature [**?**]. In a Bayesian NN, the probability distribution of network weights, conditioned on the data, is learned. Prediction of the network output can be obtained using the expectation of the posterior distribution with respect to this weight distribution [**?**]. When the number of weights is large, this can be intractable. MC dropout provides an efficient alternative to a Bayesian NN. By sampling the weights used in each minimization step from a Bernoulli distribution, the posterior distribution can be approximated. Full details and bounds on the approximation are given in [**?**]. MC dropout has been used loosely in the remote sensing community as an uncertainty measure [**?**], in part because results indicating the variance or entropy of the outputs are similar to the user's intuitive expectation of model uncertainty.

In addition to epistemic uncertainty, there is also uncertainty that cannot be accounted for by modifying the model, and is often considered as the uncertainty due to the input features. This is referred to in the literature as *aleatoric* uncertainty [**? ?**], although we note the distinction between aleatoric and epistemic uncertainty is subject to interpretation [**?**]. Here, we follow [**?**] and capture the aleatoric uncertainty by placing a probability distribution over the NN output and learning the variance of that distribution.

Research in the area of uncertainty estimates in satellite retrievals of geophysical quantities is currently led by the European Space Agency Climate Change Initiative (CCI) [**?**]. For the problem of sea ice concentration retrieval methods, this problem has been investigated within the sea ice CCI [**? ?**]. In these studies, the uncertainty of the sea ice concentration retrieval is captured by looking at the standard deviation of the tie points for ice and water respectively, where the tie points describe the typical sensor signature for the given surface type. The ice concentration estimate is determined in part by a linear interpolation using these tie points. Additionally, the uncertainty model contains a term to account for the smearing that occurs due to the difference between the instrument field of view and the interpolation grid [**? ?**]. At the present time, there are no studies of uncertainty in SAR-based ice/water retrievals, although several centers are now investigating these retrievals for operational ice monitoring (e.g., Danish Meteorological Institute, Norweigan Meteorological Institute, Canadian Ice Service). The work presented here represents a first look at the problem of uncertainty quantification in this domain using a neural network.

III. Ice/water database

To train and test our neural network, we used a previously established database that consists of co-located SAR image features extracted from RADARSAT-2 ScanSAR Wide HH and HV images, with ice/water labels collected from image analyses from the Canadian Ice Service (CIS) [**?**]. The image analyses are manual analyses of the SAR image generated by a trained ice analyst. The analyses contain information on ice concentration and ice type. For example, new ice, first-year ice and multi-year ice are commonly distinguished ice types. For the purpose of this investigation, rasterized image analyses were used to enable mapping of the ice concentration information to the SAR imagery. All SAR image features were calculated over patches consisting of 41 $\times$ 41 pixels, or 2.05 km $\times$ 2.05 km. In [**?**] the patch size of 41 $\times$ 41 pixels was chosen in order to adequately represent local image texture, and scales of various features in sea ice such as floe boundaries. At the same time, the patch size is still considerably smaller than the typical spatial resolution of Arctic sea ice modeling systems, for example, the Regional Ice and Ocean Prediction system at Environment and Climate Change Canada (ECCC) has a nominal resolution of 5 km. Samples with a land boundary within 8.05 km $\times$ 8.05 km from the patch center were not included. The features in the database that are used to discriminate ice from water are the HH-HV correlation, the difference between SAR wind speed [**?**] and the 10 m wind speed from ECCC global environment multiscale model (GEM) regional deterministic forecasts, and the standard deviation of the SAR wind speed. The ice/water labels were collected from corresponding CIS image analysis products. For training the network, only labels from pure polygons (ones that are either 100% ice or 100% open water) were used to ensure very high accuracy of the ground-truth information. Sampling of polygons with intermediate ice concentrations (between 0% and 100%) for training purposes is limited by the fact that the scale of our sample (2.05 km x 2.05 km) is much smaller than the typical scale of a CIS Image Analysis polygon ($\approx$100 km). Therefore, when a large polygon

with an intermediate ice concentration is sampled with our relatively small window, the polygon can contain regions both 0% and 100% ice for a single values of ice concentration, and it is not clear how to interpret the derived samples. Table I in [**?**] summarizes the number of samples for 0%, 100%, and 0-100% ice concentrations in both the training and testing subsets.

The database was compiled using imagery analysed by CIS operations over the years 2010 - 2016, with the year of 2013 held out as an independent test set. This enables a comprehensive test set that covers a variety of ice conditions. A plot of the geographic distribution of samples in the database is shown in Fig **??**. Given the large number of samples in the database, the data points were thinned before plotting to show the representative geographic distribution. The samples are primarily from more southern latitudes during the winter months, moving to the Canadian Arctic Archipelago and Arctic Ocean in the summer months, which reflects the coverage of CIS ice charting operations. Full details of the database and the SAR image features used are given by Komarov and Buehner [**?**].

IV. Methodology

For the ice/water classification problem we use a multilayer perceptron (MLP), which is one of the most widely used NN models. An MLP consists of three types of neuron layers: an input layer, hidden layers and an output layer. Given a training sample $(\boldsymbol{x}, y)$ where input $\boldsymbol{x}$ has n features, $\boldsymbol{x} = x_1, ..., x_n$, and y is the corresponding label (0 for water or 1 for ice), the input layer is constructed by associating one neuron to each feature. Each given input vector propagates through the network layer by layer until it reaches the output layer. At layer l each neuron j has a weight W_{ij}^l for each output i from the previous layer with m neurons, a bias b_j^l, and an activation function, ψ, to produce the output a_j^l as [**?**]

$$a_j^l = \psi\left(\sum_i^m a_i^{l-1} W_{ij}^l + b_j^l\right) \tag{1}$$

When the propagation reaches to the output layer, the output of the network, $\hat{y}$, is compared to the desired target, y, using a loss function and the error is calculated. The resulting error is propagated back through the network and the value of each weight in the network is updated. This optimization technique repeatedly performs the forward and backward propagation process followed by weight update, until a stopping criterion is satisfied. The description given here follows a single sample through the network, but for efficiency and stability of the minimization the samples are normally fed through the network in batches [**?**]. Here, initial experiments were carried out testing batch sizes of 2048 and 256, Based on time efficiency and classification accuracy, a batch size of 2048 was fixed for all the models. Parameters used in the MLP for the present study are given in Table **??**.

To investigate the impact of MLP architecture on the ice/water classification, three different network architectures were examined: a shallow network, a mid-size network, and a deeper network. All architectures were evaluated with three and four input features. The three feature network used the features described in Section **??**, while the four feature network used SAR wind speed and numerical weather prediction (NWP) wind speed as two separate features instead of using their difference. The training set was split into 70% training and 30% validation sets. The training procedure was stopped if the validation loss did not decrease after five epochs or the maximum number of epochs, which is 30, is reached. To evaluate the network with an independent data set, the trained model was used to generate ice/water labels for data of year 2013, which were then compared against ice/water labels from image analysis charts.

Table I
Parameters used in the MLP approach.

MLP parameters	Parameter value
Number of input features	3,4
Number of hidden layers	1,5,10
Number of neurons/hidden layer	10, 100, 500
Activation - inner layers	Relu
Activation - last layer	Sigmoid
Loss function	Binary cross-entropy
Batch size	2048
Optimizer	Adam optimizer [**?**]
Learning rate	0.001
Number of epochs	30

V. Uncertainty estimation

A. Epistemic uncertainty

Epistemic uncertainty is modeled by placing a prior distribution over the model weights at test time, and observing the network predictions for each set of weights drawn from the distribution. To generate different distributions of the network weights, dropout is used [**? ?**]. In this approach, for a neural network with L hidden layers, for any layer $0 \leq l \leq L$ with neurons $\boldsymbol{a}^l$, a binary vector $\boldsymbol{r}^l$ of the same size is generated by sampling from a Bernoulli distribution with probability $1-p, 0 \leq p \leq 1$. The output of layer l is multiplied element-wise with vector $\boldsymbol{r}^l$ as

$$\begin{aligned} \boldsymbol{r}^l &\sim \text{Bernoulli}(1-p), \\ \hat{\boldsymbol{a}}^l &= \boldsymbol{a}^l \odot \boldsymbol{r}^l, \end{aligned} \tag{2}$$

In this approach, p portion of outputs of layer l are set to 0, which results in thinned outputs at layer l. This thinned output provides the input for the next layer of the network. Using this approach for each given training case, a subnetwork is sampled from a larger network, denoted here as $\hat{W}_t$ for the set of weights corresponding to the $t'th$ sample of network weights. The model prediction for input feature vector $\boldsymbol{x}$ is evaluated as,

$$p = p(y=1|\boldsymbol{x}, \boldsymbol{X}, \boldsymbol{Y}) \approx \frac{1}{T}\sum_{t=1}^{T} \text{sigmoid}(f^{\hat{W}_t}), \tag{3}$$

where the total set of input features is $\boldsymbol{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_N\}$ with size N, corresponding labels $\boldsymbol{Y} = \{y_1, ..., y_N\}$, and T the number of sets of weights considered. From these

outputs (which can be considered uncalibrated probabilities) the entropy is calculated as,

$$H = -\left(p \log p + (1-p)\log(1-p)\right). \quad (4)$$

Following [**?** **?**] this entropy can be taken to be equivalent to the epistemic uncertainty. Note that entropy is a maximum for $p = 0.5$ and drops to zero for $p = 0$ and $p = 1$.

B. Aleatoric uncertainty

There are two types of aleatoric uncertainty discussed in the literature: task-dependent (or homoscedastic uncertainty) and data-dependent (or heteroscedastic uncertainty). Homoscedastic uncertainty assumes constant observation noise for different inputs of a problem while heteroscedastic uncertainty depends on the inputs of the problem and is predicted as a model output. Heteroscedastic uncertainty is very useful in remote sensing applications when data dependent observation noise is present due to measurement conditions and numerical retrieval methods.

Heteroscedastic uncertainty for a regression NN can be estimated by placing a Gaussian (or Laplacian) prior over the output from the last hidden layer. The variance is added to the output by adding a neuron to the output layer. This variance is then estimated in the minimization of the loss function [**?**]. For classification problems, Kendall and Gal proposed an approach similar to regression to calculate the heteroscedastic uncertainty [**?** **?**]. Generally, in a binary classification neural network model, for a given sample the network predicts a unary f, which is passed through a sigmoid function to produce an output p. For the case of interest here the proposed method for calculation of uncertainty places a Gaussian distribution over the unaries where the distribution variance is predicted as one of the model outputs,

$$\begin{aligned} g_i|W &\sim \mathcal{N}(f_i^W, (\sigma_i^2)^W), \\ p_i &= \text{sigmoid}(g_i). \end{aligned} \quad (5)$$

In (**??**), W represents the network parameters and f_i^W and σ_i^W are the network outputs. For each sample i, output f_i^W is perturbed by Gaussian noise with variance $(\sigma_i^2)^W$ to produce the perturbed output, g_i. The result is passed through the sigmoid function to obtain an output p_i. This is carried out T times, with each trial employing a different perturbation of the variance $(\sigma_i^W)^2$. For the binary case, the cross-entropy loss function is used, which is defined as

$$loss = -\sum_{i}^{N} \frac{1}{T} \sum_{t}^{T} \left(y_i log p_{i,t} + (1-y_i) log(1-p_{i,t})\right), \quad (6)$$

where $loss$ is the total loss function, T is the total number of Monte Carlo runs for Gaussian sampling indexed by t, N is the number of samples, and $p_{i,t}$ is the neural network output for the i^{th} sample for the t^{th} Monte Carlo run. The loss function in (**??**) ideally would be close to zero when the model predicts low noise variance σ_i and the sigmoid output yields either of its limits (zero or one), which correspond to the true classes.

C. Combined uncertainty

By combining the method of epistemic uncertainty prediction together with the heteroscedastic aleatoric uncertainty, we have a NN model that is able to learn both at the same time [**?**]. The combined model adds a dropout layer after each hidden layer and produces two outputs, one is the output of the signmoid function, which can be interpreted as an uncalibrated probability of ice or water, and the other is the variance of the unary. The combined model uses Monte Carlo (MC) twice: 1) during training when the logits are perturbed using samples drawn from a distribution that has variance equal to the aleatoric uncertainty 2) during the testing when a set of predictions is made using different sets of weights. Each set of MC runs is carried out 100 times. However, because each MC run does not correspond to a separate re-training of the model, this does not have a significant impact on the overall time required to obtain the predictions. For example, for 2.5 million test samples, the test time is longer by approximately three minutes as compared to the case without uncertainty.

VI. Description of Experiments

A. Experiments using the base MLP model

The first set of experiments carried out investigated various MLP architectures. Given that the objective was not to find the optimal MLP architecture, an exhaustive investigation was not carried out. Both the number of layers and the number of weights/layer were varied empirically. Systematic approaches are given in [**?**]. The number of weights varied between 501 for the single layer network with three input features, to 277,010 for the ten layer network with four input features. For the shallow networks there were 100 nodes in each layer, while for the deeper ones there were more nodes in the middle layers (to a maximum of 500) and fewer for the outer layers. All MLP models were implemented using the Keras API with Tensorflow as the backend [**?**].

After the MLP architecture was chosen, experiments were carried out to investigate the idea of using only one year of training data instead of all six years. This scenario is more realistic for operational implementation of the method as a long time series of training data is not generally available. The one year subset of all samples is more likely to cover the regions, seasons and weather conditions of the full training dataset compared to drawing random samples from the entire dataset. This context is also more relevant for applications where data may be collected over a previous year, but there is insufficient storage or sensor continuity for longer time periods.

B. Experiments incorporating uncertainty

The uncertainty approaches are evaluated to assess their impact on the scores, their ability to generate meaningful uncertainty maps for the ice/water classification problem in addition to insight into the input features associated with high uncertainty. To keep the problem simple, these experiments only used MLP models with one hidden layer.

For the experiments regarding the epistemic uncertainty, the dropout rate was set to 5%. This is lower than the

conventional dropout rate of 50% that is used for big and deep networks, because the networks used here are shallow and small. Since the uncertainty estimation approaches need more epochs to converge, the maximum number of training epochs were increased from 30 to 50. The early stopping criteria were kept the same as that for the experiments without uncertainty.

For the model configurations incorporating uncertainty, additional testing was carried out to evaluate model performance as a function of the training data. In addition to using only one year of training data (as was done for the base MLP model) an additional experiment was carried out in which training data was extracted over different seasons. For this experiment, training samples for the year of 2014 were divided to three subsets based on their month of acquisition. For each subset, an MLP was trained to see if it is possible to have a reasonable outcome if for any reason, only having data from a specific season is available for training. The test set is the same as before, which is the data from the entire year of 2013. By using the same test set with different training data, the impact of the training data on the scores can be more clearly delineated. For simplicity, all the models for this experiment are trained with three features.

C. Postprocessing of the MLP output and calculation of the scores

The last layer of the MLP produces an output is bounded by the sigmoid function between 0 and 1. Labeling of the samples as ice or water based on this output is done following the same approach as Komarov and Buehner [**?**]. Briefly: if the MLP output is greater than 0.95, the sample is labeled as ice and if it is less than 0.05, it is labeled as water. If it is between 0.05 and 0.95, this sample is labeled as unknown.

To calculate the scores, the ice/water labels from applying these thresholds to the MLP output need to be compared with ice/water labels from verification data. These verification data are obtained from operational ice charts. Two different methods are used. The first method uses only pure ice/water polygons from the ice charts directly as the ice/water labels. The second method is to assign every ice concentration data point from the ice charts with value less than 30% the label water, and every ice concentration data point from the ice charts with value greater than 30% the label ice. The evaluation with pure ice and water labels is carried out to assess MLP performance using a test data set that is generated in a similar manner to the training data set. It is expected that the performance of the MLP will be optimal in this scenario. The evaluation of the MLP using all ice concentration values in the ice charts is carried out to address the ability of the MLP to classify an entire SAR image, which will generally contain a range of ice concentrations. An ice concentration value of 30% is chosen as the threshold based on the study by Komarov [**?**].

For measuring the accuracy of each model, six statistical parameters are evaluated and reported: 1) training loss; 2) ice accuracy, as the percentage of correctly classified pure ice samples; 3) water accuracy, as the percentage of correctly classified pure water samples; 4) total accuracy, derived by percentage of correctly classified samples from both classes; 5) total misclassified samples; and 6) percentage of unknown samples. The number of misclassified samples is particularly important for data assimilation applications because these samples will lead to errors in the analysis, which could result in unphysical behaviour or biases when the analysis is used to initialize an ice-ocean model.

VII. Results

A. Results from the base MLP model

1) Impact of the MLP architecture: Table **??** displays the scores of the models along with their training time. With a fixed number of features, increasing the number of hidden layers in the neural network does not change the accuracy of the model significantly. When the number of hidden layers is increased, the accuracy of the MLP models trained on three features decreases slightly from 80.51% to 78.25%, for the MLP models trained with four features the accuracy increases slightly with increasing number of hidden layers, from 85.86% to a maximum of 86.20%. Comparing the three feature and four feature models, the idea of separating the NWP and SAR wind speeds is shown to increase the accuracy in terms of all measures, specifically the accuracy of water samples (6-12% improvement) which results in increasing the overall accuracy by 5%. In addition, the percentage of unknown samples decreased by almost 5% when NWP and SAR winds were included as separate features. Given that there are approximately 6M samples in the test dataset, a 5% reduction in the number of unknowns corresponds to about 120,000 samples.

The training time of each MLP model on NVIDIA Titan X GPU is also reported in Table **??**. Regardless of number of input features, networks with one hidden layer required only 6 minutes in total to train while the deep networks required 27.5 and 45.5 minutes. This means that the required training time of the deeper networks is 4.5 and 7.5 times more than shallow networks for networks with three and four input features respectively. Note that although there are a sufficient number of training samples for a deeper network, we did not investigate this because it was desired for the training time of the basic network to be less than one hour.

2) Impact of using only one year of training data: The results from training using only one year of data, which was chosen as 2014, are given in Table **??**. These models are denoted as MLP_2014. These results are broadly consistent with those from training on the full dataset. The MLP networks trained on four features have higher total accuracy compared to the three feature models. The total misclassifications and unknown samples are also similar to those from the full training data set, indicating that reasonable results can be obtained using only one year of training data. In addition, since the training set has become smaller, one epoch of the training process is much faster. As an example, the single layer MLP model requires 25 seconds to train using three features in comparison with 6 minutes to train on the whole dataset, while the two approaches have similar scores.

B. Results from incorporating uncertainty in the MLP

Table **??** shows that adding epistemic uncertainty to each MLP reduces the accuracies by 1-2%. This is mainly due to the small networks used in this study. Since the number of weights to be learned ($O \approx 100$) is significantly less than the number of samples ($O \approx 1M$), dropout reduces the network performance. Likewise, increasing the dropout ratio was observed to further decrease accuracy. The classification scores illustrate that some water samples have been identified as unknown samples while the accuracy of the ice class has not changed much. However, the fraction of misclassified ice samples decreases in all cases, and for MLP_2014 the fraction of both misclassified ice and water decreases. Fig **??** visually represents the epistemic uncertainty and MLP prediction maps of the results for the MLP_3 model when standard deviation of the SAR wind speed is 1.5. The epistemic uncertainty is higher on the decision boundary, and where that boundary extends into the unobserved region (outside of the red polygon in Fig 2a).

To investigate the impact of using different thresholds on the MLP output on the scores, in Fig **??** the accuracy, unknown percentage, and misclassification rates for ice and water are plotted versus a variation of thresholds for MLP 3 with and without epistemic uncertainty. The figure shows that adding epistemic uncertainty changes the scores of the water class more than those of the ice class. In addition, this illustrates that the choice of threshold has more influence on the water class for thresholds in the range of 3% to 20%. For thresholds less than 2% the majority of water samples will be labeled as unknown rather than correct water label.

The results of employing aleatoric uncertainty in the MLP models are also shown in Table **??**. Similar to epistemic uncertainty, adding aleatoric uncertainty slightly reduces the water accuracy and total accuracy of the MLP models and increases the unknown ratio. Despite these accuracy reductions, for each MLP category and its variations of added uncertainty, the models with added aleatoric uncertainty have lower ice misclassification rates. The distribution of the logits and their predicted variances for MLP_3 with added aleatoric uncertainty are shown in Fig **??**. The vertical lines in panel (a) of the show that the misclassification rate of ice class is higher than water class, while the distribution of variances in panel (b) shows that ice samples in the test dataset are more likely to have a slightly higher uncertainty. The map of aleatoric uncertainty (Fig **??**c) shows the aleatoric uncertainty is low in the middle of the feature space, which is where the training data are located, and increases in the unobserved regions.

The combined model estimates both the aleatoric and epistemic uncertainty. As expected, the accuracies reported in Table **??** show that the combined model modifications on the MLPs do not increase their total accuracy and the results are similar to having only either aleatoric or epistemic uncertainty. However, the total misclassification rate of MLP_2014 is the lowest of the MLP models with a score of 0.36%.

1) Evaluation using verification data from all ice concentrations: Table **??** shows impact of adding uncertainty to the MLP models when samples covering all ice concentration values are used for testing. Similar to the results on pure ice and water samples, adding uncertainty was observed to reduce the overall accuracy of their original MLP models by about 2% but the misclassified rate was also reduced in most cases, at the cost of increasing the percentage of unknown labels.

C. Results from using seasonal training data

Before presenting results from the experiment in which training was carried using seasonal data, we look at the distribution of each subset of the training data along with that from the entire year of 2014 and the entire test dataset (Fig **??**). The histograms show that except for the period of May to August, the training data sets have similar distributions of ice and water samples for each feature. Table **??** also shows the size of each subset and their results on the prediction of the test dataset.

1) Impact of seasonal training data on the scores: The scores in Table **??** show that the network trained with data from January to April (MLP_JA) is able to correctly classify ice samples better than the other models, with the highest score of accuracy and lowest score of misclassification. However, this network performs poorly on water samples, as the water class accuracy is below 50%. This could be related to the overlap for the variance of SAR wind speed training data histograms, which is greater for Jan-Apr than for other periods (Fig 6). In contrast, MLP_MA which is trained mostly on summer data, has the highest ice misclassifications, which leads to the highest total misclassification scores.

2) Relationship between input features, probabilities and uncertainties: Histograms are shown in Fig **??** that illustrate the MLP output and uncertainty for each feature input to the MLP. Separate histograms are shown for points in the test data that correspond to ice and water respectively. The results shown are based on predictions of MLP_ 2014 and MLP_ MA. These networks were chosen because MLP_ MA is a subset of MLP_ 2014, and has a different distribution of training data and reduced number of samples.

The first two columns in Fig **??** are probabilities of ice predicted by the MLP noted at the top of the column. We expect bright yellow and dark green for the histograms corresponding to ice and water samples respectively. Deviations from bright yellow and dark green indicate samples with intermediate MLP output values, which are often seen where the PDFs overlap. For example, the first two columns of the top row, which are sorted according to SAR-NWP wind speed, shows that this feature could lead to the most ice misclassifications, as it has more green colour on the left side of the ice distribution, indicating a high likelihood of water. This is more noticeable for the MLP_MA, as also numerically reported in Table **??**. These intermediate probabilities correspond to relatively large negative values of SAR-NWP windspeed (bottom axis), which might be helpful information for reducing misclassifications.

The middle two columns of Fig **??** are coloured according to the values of epistemic uncertainty. For SAR-NWP windspeed, it can be seen there is a higher uncertainty on the left side of the epistemic uncertainty histograms, which corresponds to the intermediate values. The epistemic uncertainty is also

increased for higher values of the standard deviation of SAR wind speed, although this does not appear to lead to potential misclassifications, as the corresponding ice and water histograms show high and low values respectively.

The last two columns are colored according to the values of aleatoric uncertainty. Fig **??** shows a distinct difference between the uncertainties associated with MLP2014_ and MLP_ MA, with the latter (summer acquisitions) being significantly higher than the former. This is consistent for both ice and water samples.

Fig **??** represents an example of predictions made by MLP_ 4 and MLP_ 4 with combined uncertainty for an image acquired on May 3, 2013 over the Labrador Sea. The MLP model with combined uncertainty shows a reduced number of misclassified water samples. As represented in Fig **??**, the model has difficulties in classifying water samples was due to the existence of wind over open water areas which resulted in similar feature pattern in open water areas and ice covered areas, as shown in panel (d).

VIII. Discussion

The results shown here support the use of methods presented in the literature to estimate uncertainty from NNs for the problem of retrieval of ice and water from SAR imagery. Referring back to the questions posed in the introduction, we first compare the MLP results to those from logistic regression [**?** **?**], which are also shown in Table **??**. The first logistic regression method [**?**] can be compared with the three feature version of MLP_ALL, as it is using the same three features as input. From Table II it can be seen the MLP has a slightly lower ice accuracy and higher ice misclassification rate than the logistic regression, while the water accuracy and misclassification rates are improved when the MLP is used. Overall, when an MLP with 10 hidden layers is used the total misclassification rate is comparable to that from logistic regression, with a higher accuracy and fewer unknowns. It is noteworthy that Komarov et al. improved upon their first logistic regression classifier by developing a method to adapt the thresholds (instead of fixed thresholds of 0.95 and 0.05) to classify the logistic regression output as ice and water [**?**]. This method is more comparable to the four feature model because it takes into account the SAR wind speed as an independent variable to adapt the threshold. It can be compared to results from the MLP models if a more stringent threshold is used to classify ice and water for MLP models. For example, using thresholds of 0.98 and 0.02 to identify ice and water for the four feature MLP model leads to a reduced misclassification rate (0.25%) and a higher fraction of unknowns (26.18%), similar to logistic regression with adaptive thresholds. The model used in practice should be chosen based on the requirements of the application.

The proposed uncertainty estimation methods do not improve the accuracy of the MLP models, but they do result in lower misclassification rates. The epistemic uncertainties are higher along the decision boundary. This is expected because the MLP output (uncalibrated probability) takes on intermediate values along the decision boundary, and the entropy is high for intermediate values of the output (reaching a maximum at $p = 0.5$). The map of aleatoric uncertainty in feature space, Fig **??**, shows that the aleatoric uncertainty values increase when moving away from the part of the feature space with training data, which is consistent with the noted seasonal dependence. Overall, the values for aleatoric uncertainty are low. The values given in Fig **??** and Fig **??** are the log of the variance of the aleatoric uncertainty. For example, a value of -5 corresponds to a standard deviation (in logit space) of 0.082, and a value of -20 corresponds to a standard deviation of 4.5×10^{-5}. These are the perturbations that one would expect to be applied to the logit before it is input to the sigmoid function. While these values are small, and one might wonder how the overall output could be impacted by such a small value, we note that the variances are obtained in tandem with the MLP weights during the optimization. Hence, changes in the output are not due to the variance alone, but also due to the fact that because of the presence of the variance, the learned weights will differ between the model with and without the aleatoric uncertainty. This should be investigated further in a future study. At the present, we note that similarly small values of aleatoric uncertainty are found in the study by [**?**].

Regarding the relationship to input features, the epistemic uncertainty is slightly enhanced for both the ice and water classes for values of SAR-NWP wind speed for which the two MLP output (uncalibrated probability) distributions (the one for ice and the one for water) overlap. These points also correspond to intermediate output values. The epistemic uncertainty is also enhanced for the water class for higher values of the standard deviation of the SAR wind speed (bottom row of Fig 6). For the aleatoric uncertainty, the noteworthy link to input features is through the seasonal dependence. For all features, the aleatoric uncertainty is significantly enhanced for the model trained on data from May to August of 2014 and tested on the full year of test data. For the model trained on a full year of data, the aleatoric uncertainties are small.

IX. Conclusions

The present study has shown that the MLP models without uncertainty are able to classify the SAR imagery into ice and water with accuracies that are comparable to those from the logistic regression. The fact that the MLP models are able to classify SAR imagery into ice and water given the database from [**?**] shows that these features are indeed very useful for ice/water discrimination, and that the high classification accuracies in [**?**] are not necessarily a function of the classifier used. While a logistic regression may be more robust than a NN, we do think the non-linearity of the decision boundary may be better captured in the NN, leading for some cases to fewer misclassifications. The MLP models investigated here always exhibited a smooth decrease in the loss function for both the training and validation data, indicating that they are well-behaved, which may be due to the small number of weights relative to the size of the training database, with the number of weights typically $O(1000)$ and the number of samples $O(1M)$.When uncertainty is included, the misclassification rates predicted by the MLP models are reduced, and the

fraction of samples that are unknown is also slightly increased. An examination of the relationship between the input features and the estimated MLP uncertainties and probabilities shows that the difference between the SAR and NWP wind speed may be contributing to the ice misclassification rate. Looking at the uncertainties and probabilities for a given day off the Labrador coast demonstrates that the uncertainties may be useful to flag regions in the MLP predictions that should be checked manually by an analyst. The lower misclassification rates that are achieved when uncertainty is included in the model also suggests that the ice and water classifications be useful for data assimilation. Future work should investigate the use of these uncertainty estimates as well as the impact of the source of training data labels and verification data. Image analysis charts share similar characteristics to other manual analyses, and may have biases and errors, particularly for intermediate ice concentrations [**?**].

Nazanin Asadi received the B.Sc. and M.Sc. degrees in computer engineering from the Isfahan University of Technology, Isfahan, Iran, in 2010 and 2012, respectively, and the Ph.D. degree in systems design engineering from the University of Waterloo, Waterloo, Canada, in 2019. Her current research interests are applied machine learning, remote sensing and data assimilation.

K. Andrea Scott received the B.A.Sc. and Ph.D. degrees from the University of Waterloo, Waterloo, ON, Canada, in 1999 and 2008, respectively, and the M.A.Sc. degree from McMaster University, Hamilton, ON, in 2001. She was a Post-Doctoral Researcher with the Data Assimilation and Satellite Meteorology Research Section, Environment and Climate Change Canada, Toronto, ON, 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.

Alexander S. Komarov (S10/M15) received the B.Sc. (Hons.) degree in Radiophysics and Electronics and the M.Sc. (Hons.) degree in Physics from the Altai State University, Russia, in 2006 and 2008, respectively, and the Ph.D. degree in Electrical Engineering from the University of Manitoba, Winnipeg, Canada in 2015. He has conducted postdoctoral research at the Centre for Earth Observation Science, University of Manitoba. Since 2015 he has been a Research Scientist with the Meteorological Research Division, Environment and Climate Change Canada. His current research interests include SAR remote sensing of sea ice and the ocean surface, SAR data assimilation, and electromagnetic wave scattering from sea ice.

Mark Buehner received the B.A.Sc. degree from the University of Waterloo, Waterloo, ON, Canada, in 1994, and the Ph.D. degree in physical oceanography from Dalhousie University, Halifax, NS, Canada, in 2000. He is currently the Scientific Lead for the development of sea ice data assimilation systems with Environment and Climate Change Canada, Dorval, QC, Canada, where he is a Senior Research Scientist with the Data Assimilation and Satellite Meteorology Research Section. His research focuses on data assimilation for both weather prediction and sea ice applications.

David A. Clausi (S93/M96/SM03) received the B.A.Sc., M.A.Sc., and Ph.D. degrees in systems design engineering from the University of Waterloo, Waterloo, ON, Canada, in 1990, 1992, and 1996, respectively. After completing his Ph.D., he worked in the medical imaging field with Mitra Imaging Inc., Waterloo. He started his academic career as an Assistant Professor in geomatics engineering with the University of Calgary, Canada, in 1997. He returned to his alma mater in 1999 and was awarded tenure and promotion to an Associate Professor in 2003. He is an active interdisciplinary and multidisciplinary researcher. He has an extensive publication record, publishing refereed journal and conference papers on remote sensing, computer vision, algorithm design, and biomechanics. His primary research interests include the automated interpretation of synthetic aperture radar sea ice imagery, in support of operational activities of the Canadian Ice Service. The research results have led to successful commercial implementations. Dr. Clausi is a recipient of numerous scholarships, conference paper awards, and two Teaching Excellence Awards.

REFERENCES

[1] C.M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[2] M. Bishop, Christopher. *Pattern recognition and machine learning*. Springer, Science and Business Media, LLC, 2006.

[3] François Chollet et al. Keras. https://keras.io, 2015.

[4] A. Der Kiureghian and O. Ditlevsen. Aleatoric or epistemic? Does it matter? Proceedings of the Special Workshop on Risk Acceptance and Risk Commmunication, March 26-27, Stanford University, 2007.

[5] Y. Gal and Z. Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 1050–1059, New York, USA, 2016.

[6] T. Geldsetzer and J.J. Yackel. Sea ice type and open water discrimination using dual co-polarized C-band SAR. *Canadian Journal of Remote Sensing*, 35:73–84, 2009.

[7] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[8] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. arXiv:abs/1207.0580, 2012.

[9] Karvonen J., M. Simila, and M. Mäkynen. Open water detection from Baltic sea ice Radarsat-1 SAR imagery. *IEEE Geoscience and Remote Sensing Letters*, 2(3):275–279, 2005.

[10] M. Kampffmeyer, A-B. Salberg, and R. Jenssen. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016.

[11] J. Karvonen. A sea ice concentration estimation algorithm utilizing radiometer and SAR data. *The Cryosphere*, 8(5):1639–1650, 2014.

[12] J. Karvonen, J. Vainio, M. Marnela, P. Eriksson, and T. Niskanen. A comparison between high-resolution EO-based and ice analyst-assigned sea ice concentrations. *Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(4):1799–1807, 2015.

[13] A. Kendall. *Geometry and uncertainty in deep learning for computer vision*. PhD thesis, University of Cambridge, 2018.

[14] A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584, 2017.

[15] D.P. Kingma and J.L. Ba. ADAM: A method for stochastic optimization. In *Proceedings of the 5th International Conference on Learning Representations(ICLR2017)*, 2017.

[16] A.S. Komarov and M. Buehner. Automated detection of ice and open water from dual-polarization RADARSAT-2 images for data assimilation. *IEEE Transactions on Geoscience and Remote Sensing*, 55:5755–5769, 2017.

[17] A.S. Komarov and M. Buehner. Adaptive probability thresholding in automated ice and open water detection from RADARSAT-2 images. *IEEE Geoscience and Remote Sensing Letters*, 15(4):552–556, 2018.

[18] A.S. Komarov, V. Zabeline, and D.G. Barber. Ocean surface wind speed retrieval from C-band SAR images without wind direction input. *IEEE Transactions on Geoscience and Remote Sensing*, 52(2):980–990, 2014.

[19] Y. Kwon, J-H. Won, B.J. Kim, and M.C. Paik. Uncertainty quantification using Bayesian neural networks in classification: Application to ischemic stroke lesion segmentation. In *1st Conference on Medical Imaging with Deep Learning (MIDL)*, 2018.

[20] T. Lavergne, A.M. Sorensen, S. Kern, R. Tonboe, D. Notz, A. Signe, L. Bell, G. Dybkjaer, S. Eastwood, C. Gabarro, G. Heyster, M.A. Killie, M.B. Kreiner, J. Lavelle, R. Saldo, S. Sandven, and L.T. Pedersen. Version 2 of the EUMETSAT OSI SAF and ESA CCI sea ice concentration climate data records. *The Cryosphere*, 13:49–78, 2019.

[21] Q.V. Le, A.J. Smola, and S. Canu. Heteroscedastic Gaussian process regression. In *Proceedings of the 22nd international conference on Machine learning*, pages 489–496. ACM, 2005.

[22] S. Leigh, Z. Wang, and D. A. Clausi. Automated ice-water classification using dual polarization SAR satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 52(9):5529–5539, 2014.

[23] C.J. Merchant, F. Paul, T. Popp, M. Ablain, S. Bontemps, P. Defourny, R. Hollmann, T. Lavergne, A. Laeng, G. deLeeuw, J. Mittaz, C. Poulsen, A.C. Povey, M. Reuter, S. Sathyendranath, S. Sandven, V.F. Sofieva, and W. Wagner. Uncertainty information in climate data records from Earth observation. *Earth System Science Data*, 9:511–527, 2017.

[24] M.A.N. Moen, A.P. Doulgeris, S.N. Anifinsen, A.H.H. Renner, N. Hughes, S. Gerland, and T. Eltoft. Comparison of feature based segmentation of full polarimetric SAR satellite sea ice images with manually drawn ice charts. *The Cryosphere*, 7:1693–1703, 2013.

[25] K.G. Sheela and S.N. Deepa. Review on methods to fix number of hidden neurons in neural networks. *Mathematical problems in Engineering*, 2013, 2013.

[26] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[27] R.T. Tonboe, S. Eastwood, T. Lavergne, A.M. Sørensen, N. Rathmann, G. Dybkjaer, L.T. Pedersen, J.L. Høyer, and S. Kern. The EUMETSAT sea ice concentration climate data record. *The Cryosphere*, 10:2275–2290, 2016.

[28] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, and T. Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. 2019.

[29] Q. Yu and D.A. Clausi. IRGS: Image segmentation using

edge penalties and region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30:2126–2139, 2008.

[30] N. Zakhvatkina, A. Korovsov, S. Muchenhuber, S. Sandven, and M. Babiker. Operational algorithm for ice-water classification on dual-polarized RADARSAT-2 images. *The Cryosphere*, 11:33–46, 2017.

[31] N. Zakhvatkina, V. Smirnov, and I. Bychkova. Satellite SAR data-based sea ice classification An overview. *Geosciences*, 9:152, 2019.

[32] Natalia Yu Zakhvatkina, Vitaly Yu Alexandrov, Ola M Johannessen, Stein Sandven, and Ivan Ye Frolov. Classification of sea ice types in envisat synthetic aperture radar images. *IEEE Transactions on Geoscience and Remote Sensing*, 51(5):2587–2600, 2013.

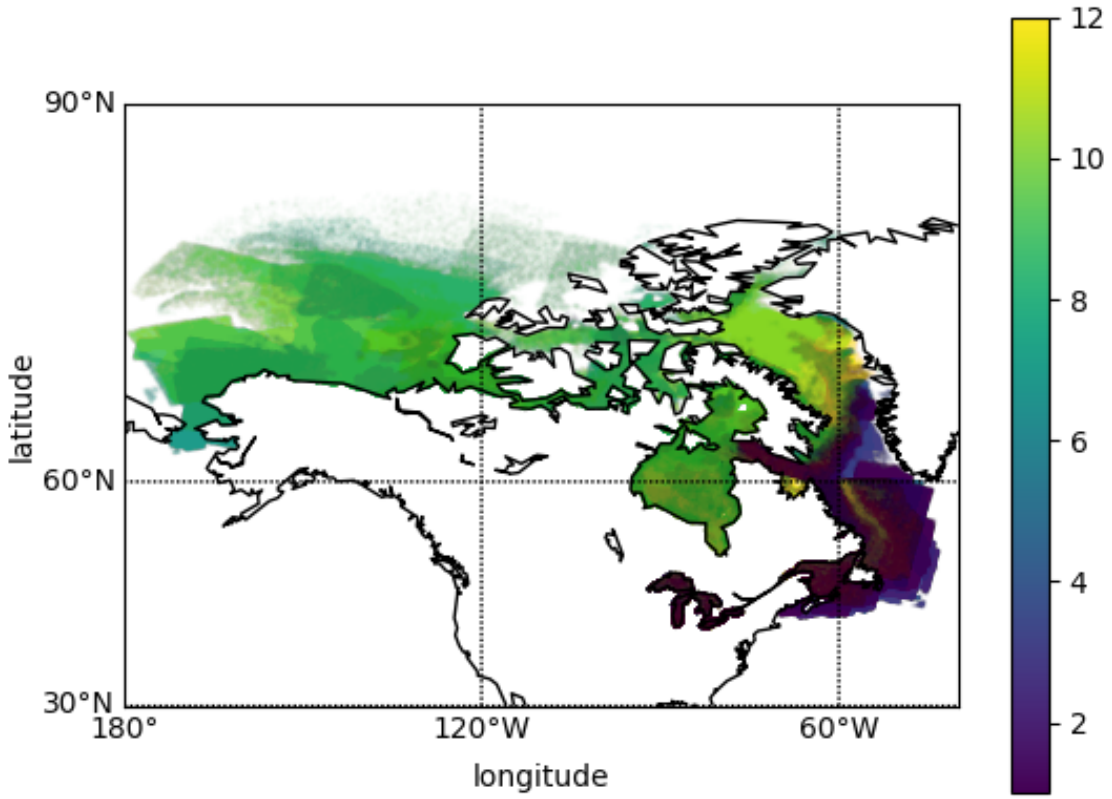

Figure 1. Map of region over which the SAR database samples were acquired. Colors indicate the month of acquisition of the SAR imagery, with 1-12 representing the months sequentially from January to December. Due to the large number of images, samples were thinned for clarity. The distribution of samples indicates that in the winter months most images were acquired on the east coast of Canada, whereas in the summer they acquired at higher latitudes. The SAR images are from the archives of the Canadian Ice Service, and this geographic distribution is representative of their ice charting practice.

Table II
ACCURACY OF THE MODELS TRAINED AND TESTED ON PURE ICE AND WATER SAMPLES. MLP_2014 MODEL IS TRAINED ON 2014 TRAINING DATA WHILE OTHER MODELS ARE TRAINED ON ALL YEARS OF TRAINING DATA.

Method	Number of features	Number of hidden layers	Training time [min]	Ice accuracy [%]	Ice misclassified [%]	Water accuracy [%]	Water misclassified [%]	Total accuracy [%]	Total misclassified [%]	Unknowns [%]
MLP_ALL	3	1	6.0	85.37	1.49	77.28	0.17	80.51	0.70	18.79
		5	10.2	86.24	1.42	76.21	0.20	80.22	0.69	19.10
		10	27.5	86.22	1.25	72.95	0.20	78.25	0.62	21.13
	4	1	6.0	89.05	1.08	83.74	0.17	85.86	0.53	13.61
		5	15.5	90.17	1.03	83.03	0.22	85.88	0.54	13.58
		10	45.5	89.40	1.10	84.07	0.19	86.20	0.55	13.25
MLP_2014	3	1	0.4	83.29	1.71	78.87	0.14	80.63	0.77	18.60
		5	1.3	82.65	1.96	82.37	0.14	82.48	0.87	16.65
		10	4.5	83.71	1.58	76.88	0.17	79.61	0.73	19.66
	4	1	1.5	88.33	0.97	80.92	0.19	83.88	0.50	15.61
		5	5.0	91.43	0.65	73.78	0.41	80.83	0.51	18.66
		10	9.0	89.35	1.12	83.77	0.29	86.00	0.62	13.77
Logistic regression [**?**]	3	-	-	88.23	0.98	61.48	0.35	72.14	0.60	27.25
Logistic regression [**?**]	Adaptive	-	-	81.35	0.24	54.85	0.09	65.44	0.15	34.42

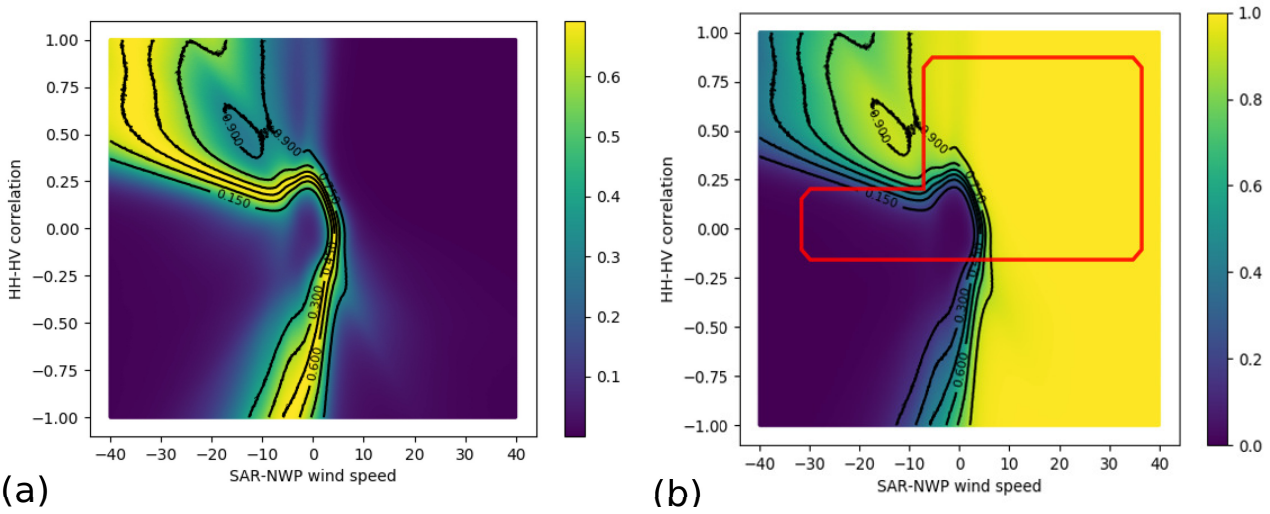

Figure 2. The MLP output (a) and epistemic uncertainty (b) map of the MLP_3 model trained with epistemic uncertainty when the standard deviation of SAR wind speed is 1.5. The contour lines of predicted probabilities are overlaid on panel (a) and (b). The model uncertainty is higher on the decision boundary and where this boundary extends into the unobserved region (a contour indicating the approximate observed region is shown in panel (a) in red).

Table III
ACCURACY OF THE MODELS TRAINED AND TESTED ON PURE ICE AND WATER SAMPLES. MODELING UNCERTAINTY IN THE MLP MODELS GIVES A SLIGHT DEGRADATION IN TOTAL ACCURACY WHILE NOTABLY DECREASING THE PERCENTAGE OF MISCLASSIFIED SAMPLES.

Method	Number of features	Number of hidden layers	Ice accuracy [%]	Ice misclassified [%]	Water accuracy [%]	Water misclassified [%]	Total accuracy [%]	Total misclassified [%]	Unknowns [%]
MLP_3	3	1	85.37	1.49	77.28	0.17	80.51	0.70	18.79
+Epistemic uncertainty			86.66	1.24	73.36	0.23	78.67	0.63	20.69
+Aleatoric uncertainty			86.80	1.20	72.65	0.22	78.30	0.61	21.09
+Epistemic & Aleatoric uncertainty			85.88	1.38	75.89	0.20	79.88	0.67	19.51
MLP_4	4	1	89.05	1.08	83.74	0.17	85.86	0.53	13.61
+Epistemic uncertainty			88.83	0.92	81.20	0.17	84.25	0.47	15.32
+Aleatoric uncertainty			89.64	0.90	80.73	0.19	84.29	0.47	15.24
+Epistemic & aleatoric uncertainty			88.73	0.92	80.96	0.17	84.06	0.47	15.48
MLP_2014	4	1	88.33	0.97	80.92	0.19	83.88	0.50	15.61
+Epistemic uncertainty			88.16	0.82	78.55	0.15	82.39	0.42	17.18
+Aleatoric uncertainty			89.65	0.77	76.94	0.24	82.02	0.45	17.53
+Epistemic & aleatoric uncertainty			89.21	0.58	72.47	0.20	79.15	0.36	20.49

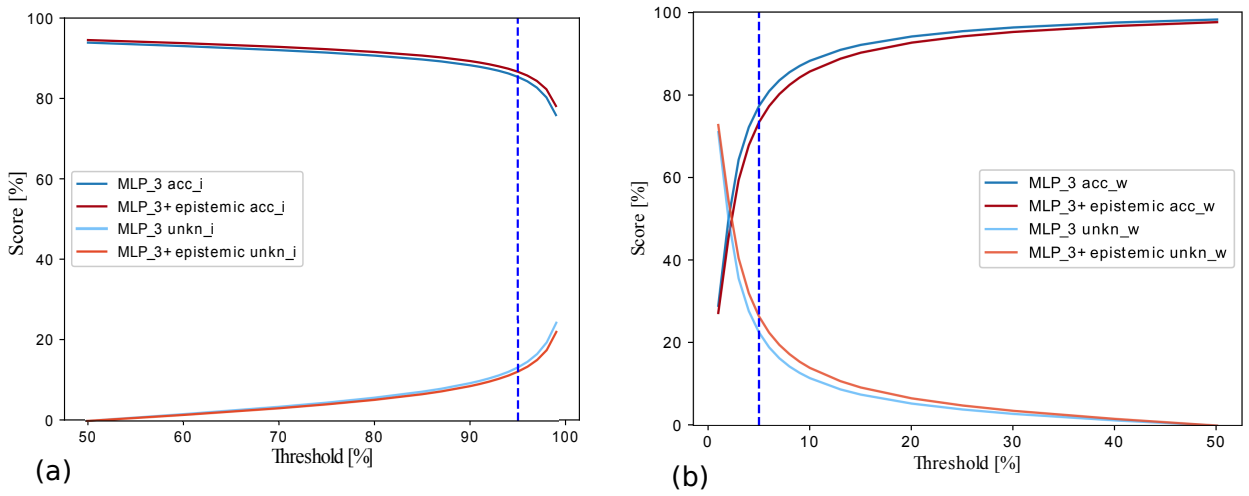

Figure 3. (a) Ice class (b) Water class. The impact of thresholds used to define ice and water points on the accuracy of ice and water classes. The selected models are MLP with 3 features with and without epistemic uncertainty. The 0.05 and 0.95 thresholds for ice and water is shown by a vertical line in each case. This figure shows that the water class is more affected by including the uncertainty.

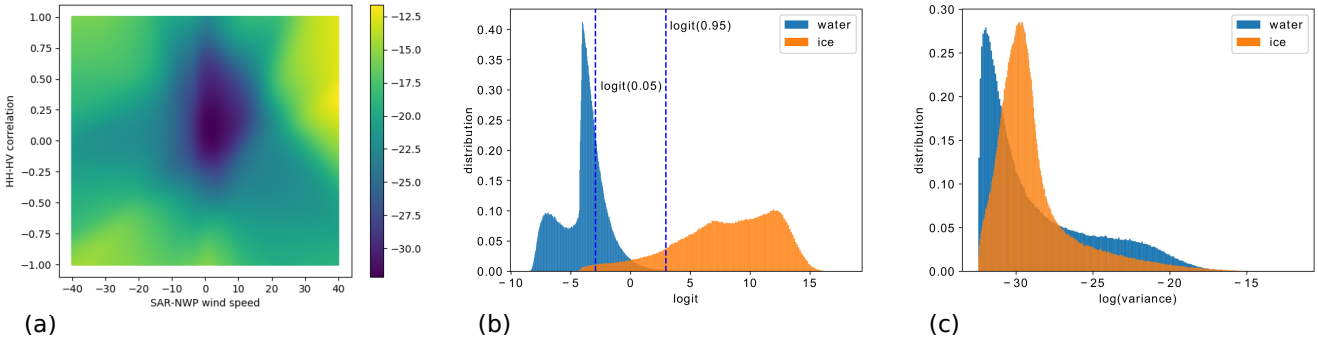

Figure 4. The aleatoric uncertainty (a) of the MLP_3 model when the standard deviation of SAR wind speed is 1.5. In panel (b) the distribution of predicted logits and variances (c) are also shown for the same model, MLP_3 including aleatoric uncertainty. The classification boundary of each class, corresponding to threshold values of 0.05 and 0.95, are indicated with vertical lines in panel (b). Note that the aleatoric uncertainty increases moving away from the center of the plot. Since the training data is more concentrated in the center of the plot, this indicates the aleatoric uncertainty increases towards unobserved regions.

Table IV
ACCURACY OF THE MODELS TRAINED ON PURE ICE AND WATER SAMPLES AND EVALUATED USING THE ALL ICE CONCENTRATION TEST SET. RESULTS SHOW THE IMPACT OF INCLUDING UNCERTAINTY IN THE MODEL.

Method	Number of features	Number of hidden layers	Ice accuracy [%]	Ice misclassified [%]	Water accuracy [%]	Water misclassified [%]	Total accuracy [%]	Total misclassified [%]	Unknowns [%]
MLP_3	3	1	78.93	4.38	75.03	2.22	76.08	2.81	21.10
+Epistemic uncertainty			79.73	4.05	72.64	2.33	74.56	2.79	22.64
+Aleatoric uncertainty			80.41	3.92	71.67	2.38	74.04	2.80	23.16
+Epistemic & Aleatoric uncertainty			79.20	4.24	74.35	2.28	75.67	2.81	21.53
MLP_4	4	1	81.70	4.61	82.54	2.09	82.31	2.77	14.92
+Epistemic uncertainty			80.18	4.21	79.95	1.88	80.01	2.51	17.48
+Aleatoric uncertainty			81.64	4.12	79.68	2.04	80.21	2.60	17.18
+Epistemic & aleatoric uncertainty			79.93	4.13	79.68	1.86	79.75	2.47	17.77
MLP_2014	4	1	77.81	5.58	80.43	1.30	79.68	2.53	17.79
+Epistemic uncertainty			79.36	3.86	77.70	1.81	78.15	2.37	19.48
+Aleatoric uncertainty			82.03	3.64	76.45	2.16	77.96	2.56	19.47
+Epistemic & aleatoric uncertainty			80.77	3.38	73.05	1.99	75.14	2.37	22.48
Logistic Regression	3	-	79.41	4.60	65.72	1.90	69.65	2.67	27.68

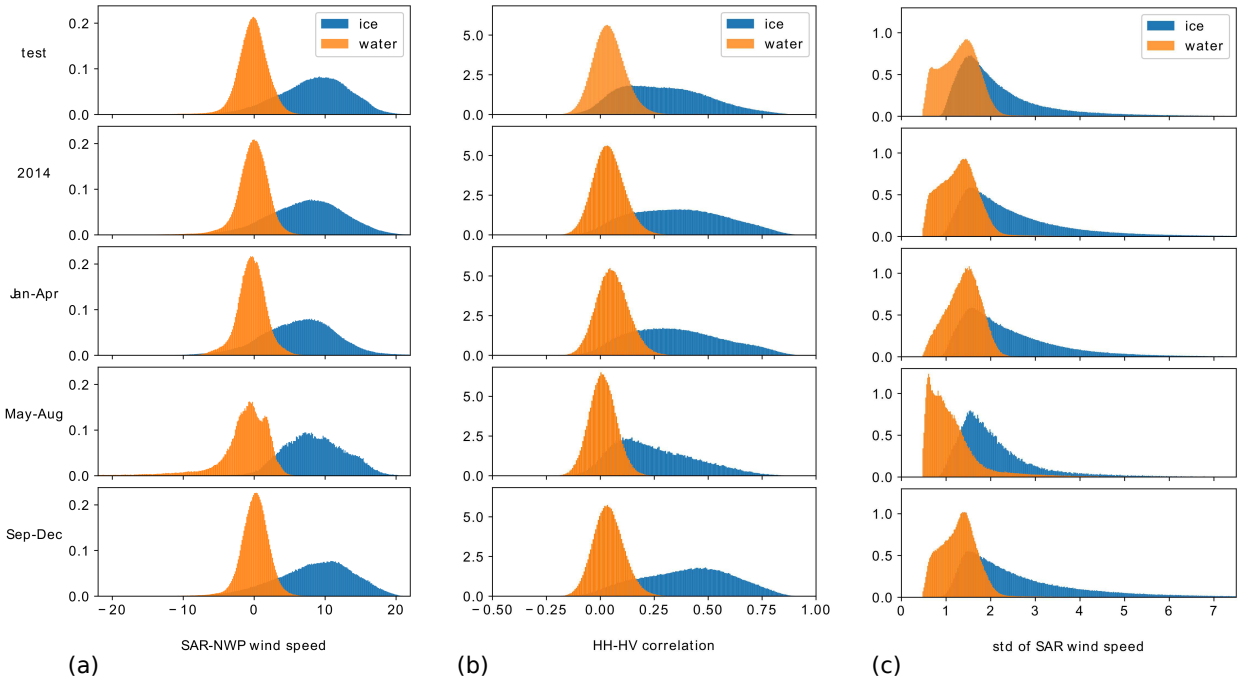

Figure 5. Distribution of ice and water samples for each of the 3 input features for (top row) the test dataset; (second row) the entire year of 2014; (rows 3-5) the three subsets of 2014.

Table V
ACCURACY OF THE MODELS TRAINED USING PURE AND ICE WATER SAMPLES FROM 2014. THE TRAINING DATA ARE SPLIT INTO THREE SUBSETS, EACH COVERING FOUR MONTHS. THE MODELS ARE TESTED ON ALL PURE ICE AND WATER SAMPLES FROM 2013.

Method	Training size	Ice accuracy [%]	Ice misclassified [%]	Water accuracy [%]	Water misclassified [%]	Total accuracy [%]	Total misclassified [%]	Unknowns [%]
MLP_2014	2,693,263	83.29	1.71	78.87	0.14	80.63	0.77	18.60
+Epistemic uncertainty		85.19	1.10	63.38	0.19	75.09	0.56	24.35
+Aleatoric uncertainty		88.33	1.06	68.43	0.22	75.58	0.55	23.87
MLP_JA	1,091,675	89.27	0.44	46.82	0.55	63.77	0.50	35.72
+Epistemic uncertainty		87.97	0.47	49.19	0.40	64.67	0.43	34.89
+Aleatoric uncertainty		89.10	0.41	45.78	0.53	63.08	0.49	36.43
MLP_MA	319,168	86.20	2.29	77.18	0.44	80.78	1.18	18.03
+Epistemic uncertainty		86.01	2.20	77.95	0.41	81.17	1.12	17.71
+Aleatoric uncertainty		87.55	2.09	74.25	0.65	79.56	1.21	19.22
MLP_SD	1,282,420	82.66	1.83	79.47	0.12	80.74	0.80	18.45
+Epistemic uncertainty		81.41	1.87	79.87	0.10	80.48	0.81	18.71
+Aleatoric uncertainty		81.69	1.78	77.95	0.11	79.44	0.77	19.78

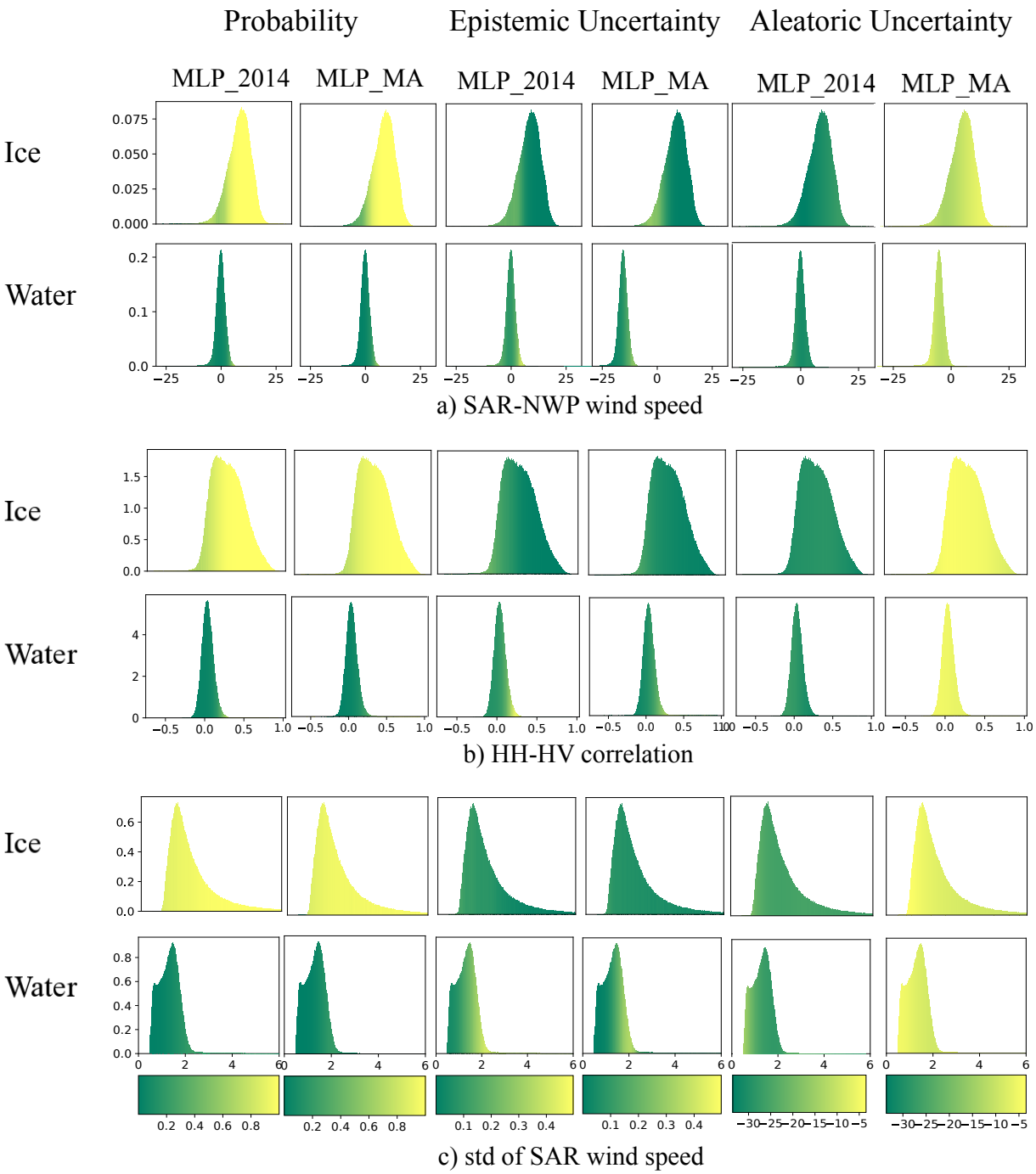

Figure 6. Impact of each feature on the MLP output and uncertainty of the MLP_2014 and MLP_MA trained on 3 features. The uncertainties given for the aleatoric case are the log of the variance. The results shown are those corresponding to the models used for epistemic uncertainty. However, results for the models corresponding to aleatoric uncertainty were visually very similar and are not shown here. The first two columns indicate the MLP output (uncalibrated probability), the middle two indicate the epistemic uncertainty, and the last two indicate the aleatoric uncertainty. The colorbars at the bottom indicate the MLP output, epistemic uncertainty and aleatoric uncertainty for each set of two columns. For the first two columns, for the rows labelled as ice, a green color indicates misclassifications, whereas for the rows labeled as water, a yellow color indicates misclassifications.

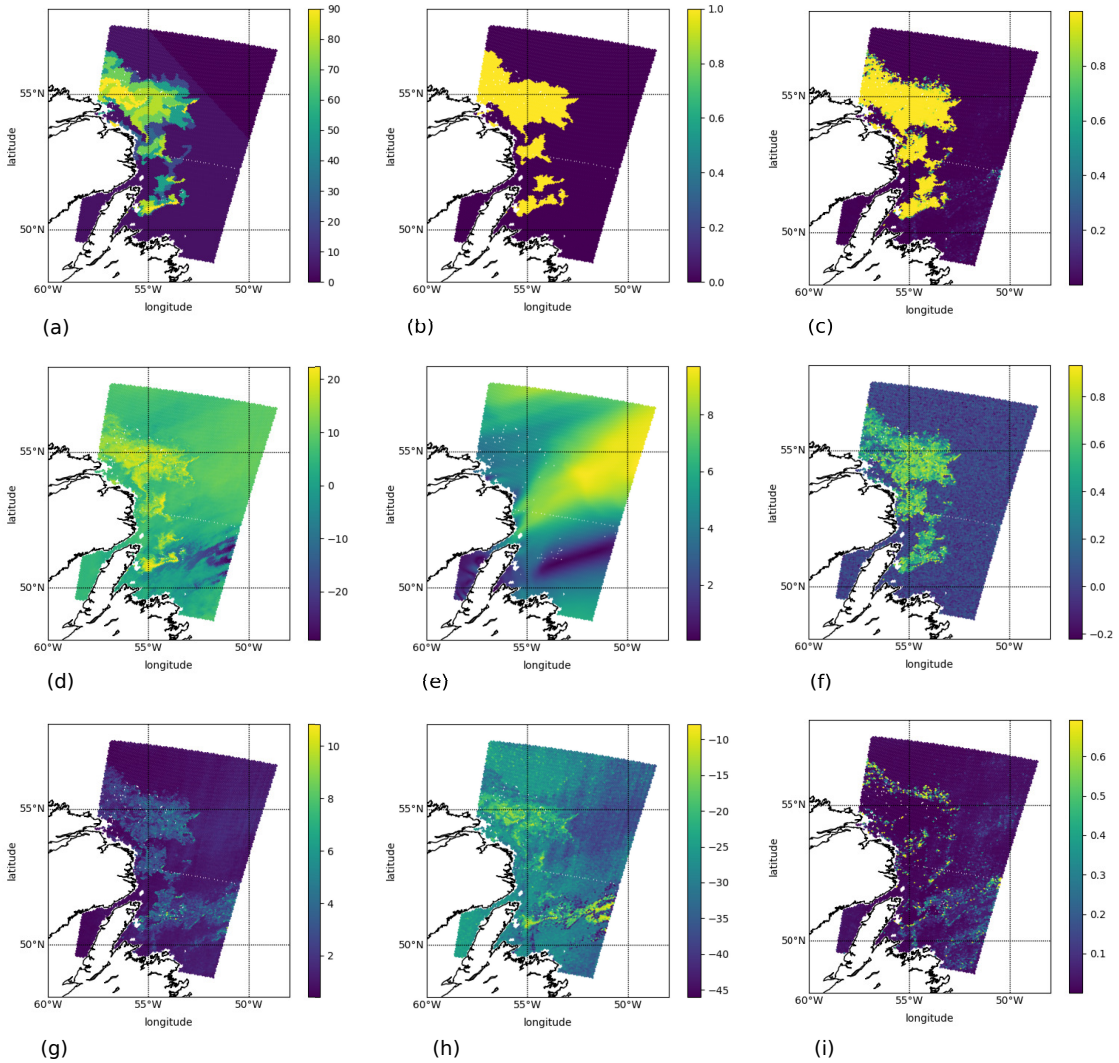

Figure 7. An example of ground truth, estimated probabilities, feature maps and uncertainty maps for an image acquired on May 3, 2013, over the Labrador Sea. Results shown are from the MLP_4 model with combined uncertainty. Aleatoric uncertainty is shown in logarithmic scale. It can be seen that the aleatoric uncertainty is high where the SAR wind speed takes on negative values, and the epistemic uncertainty is increased near the ice edge. Panels are: (a) Ice concentration (IC) (b) Ice/water labels from IC (c) MLP output (uncalibrated probability) from MLP_4 d) SAR wind speed (e) NWP wind speed (f) HH-HV correlation (g) std of SAR wind speed (h) aleatoric uncertainty (i) epistemic uncertainty