Uncertainty Estimation of Lake Ice Cover Maps From a Random Forest Classifier Using MODIS TOA Reflectance Data

Nastaran Saberi, *Member, IEEE*, Mohammad Hossein Shaker, Claude R. Duguay, *Member, IEEE*, K Andrea Scott, *Member, IEEE*, and Eyke Hüllermeier, *Senior Member, IEEE*

***Abstract*—This article presents a method to improve the usability of lake ice cover (LIC) maps generated from moderate resolution imaging spectroradiometer (MODIS) top-of-atmosphere reflectance data by providing estimates of aleatoric and epistemic uncertainty. We used a random forest (RF) classifier, which has been shown to have superior performance in classifying lake ice, open water, and clouds, to generate daily LIC maps with inherent (aleatoric) and model (epistemic) uncertainties. RF allows for the learning of different hypotheses (trees), producing diverse predictions that can be utilized to quantify aleatoric and epistemic uncertainty. We use a decomposition of Shannon entropy to quantify these uncertainties and apply pixel-based uncertainty estimation. Our results show that using uncertainty values to reject the classification of uncertain pixels significantly improves recall and precision. The method presented herein is under consideration for integration into the processing chain implemented for the production of daily LIC maps as part of the European Space Agency's Climate Change Initiative (CCI+) Lakes project.**

***Index Terms*—Lake ice, random forest (RF), remote sensing, uncertainty.**

I. Introduction

ACCORDING to the global climate observing system (GCOS), lake ice cover (LIC) is created as a thematic product of lakes as an essential climate variable (ECV) required for climate monitoring [1]. It is also a significant product of interest for improving numerical weather forecasting in northern high latitudes [2]. In recent years, a considerable number of studies have been conducted on the use of satellite-derived LIC and ice phenology (dates associated with freeze-up and breakup, and ice cover duration) records for documenting the response of northern lakes as well as lakes on the Tibetan Plateau to climate variability [3] and change [4], [5], [6], [7]. Observations from both active microwave and passive (optical and microwave) sensors have been used to map and monitor LIC (e.g., [8]). While synthetic aperture radar (SAR) sensors provide all-weather and day/night acquisitions, optical sensors collecting data in the visible to thermal infrared parts of the electromagnetic spectrum are the main instruments for environmental monitoring, especially in the context of climate change studies, which require extensive time series with global coverage. The moderate resolution imaging spectroradiometer (MODIS) aboard both NASA's Terra and Aqua satellite missions is especially relevant for this purpose, as it has been providing more than twice daily acquisitions at northern latitudes for over 20 years.

Together with the availability of a longer time series of MODIS data, there has also been an increase in the application of machine learning (ML) for lake ice classification from optical imagery [1], [2], [9], [10]. This shift of attention toward the use of ML techniques is primarily supported by the better performance in accuracy metrics compared to previous threshold-based approaches for lake ice mapping [10]. While meaningful uncertainties can enhance the explainibility of ML predictions, there is generally a gap in providing uncertainties from ML-generated outputs. Such is the case for the LIC thematic product generated for the European Space Agency Climate Change Initiative (ESA CCI) Lakes project (https://climate.esa.int/en/projects/lakes/) that is delivering multidecadal satellite-based products (lake surface water temperature, ice cover, water-leaving reflectance, water level, and extent) on a common ca. 1-km grid over more than 2000 lakes for climate monitoring and to serve the climate modeling community [11]. The LIC product delivered to the ESA CCI Lakes project currently only contains information on overall classification accuracies obtained with a random forest (RF) classifier for its main categories (open water, ice cover, and cloud cover). Pixel-based quantification of uncertainty is much needed for users of the product.

Recently, uncertainty estimation has been found an essential add-on in different ML research fields such as computer vision, and natural language processing (NLP), as well as classic

Received 8 May 2024; revised 30 September 2024; accepted 15 November 2024. Date of publication 23 December 2024; date of current version 21 February 2025. This research was undertaken thanks to funding from the Canada First Research Excellence Fund Global Water Futures Project from August 2021 to November 2022. We would like to thank Yuhao Wang for his help with data annotations for training the Random Forest classification model through support from the European Space Agency's Climate Change Initiative - New ECVs for Lakes (grant reference 4000125030/18/I-NB). *(Corresponding author: Nastaran Saberi.)*

Nastaran Saberi and Claude R. Duguay are with the Department of Geography and Environmental Managment, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: nsaberi@uwaterloo.ca; crduguay@uwaterloo.ca).

Mohammad Hossein Shaker and Eyke Hüllermeier are with the Institute of Informatics, LMU Munich, and Munich Center for Machine Learning (MCML), 80335 Munchen, Germany (e-mail: hossein.shaker@ifi.lmu.de; eyke@lmu.de).

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

Digital Object Identifier 10.1109/JSTARS.2024.3518306

ML problems such as regression and classification [12]. In satellite remote sensing, there is often an insufficient emphasis on the quantification of uncertainties associated with the derived products or maps. While these uncertainties are significant for the reliability and application of such products, there is limited literature that addresses this challenge. Notably, there are a few pioneering works in ice mapping using SAR that focus on uncertainty decomposition [13] as well as uncertainty estimation using probabilistic approaches [14]. This is particularly true when an ML approach is applied for classifying the images, whereby only basic accuracy metrics are calculated (e.g., overall classification accuracy, errors of omission, and commission) and not uncertainties. This includes the significance of differentiating two types of uncertainties often referred to as aleatoric and epistemic uncertainty. While aleatoric uncertainty refers to the randomness in the data-generating process, epistemic uncertainty is caused by the learner's lack of knowledge about the best prediction [12]. Thus, the latter could in principle be reduced through further information (e.g., more training data), whereas the former is irreducible and implies an unavoidable prediction error. The primary motivation for decomposing uncertainty into epistemic and aleatoric components is to enhance our understanding of model performance and the inherent randomness within the data. This, in turn, facilitates the planning for the fusion of various models, aiming to create a reliable lake ice map product.

In this article, we present an approach to measure the aleatoric and epistemic uncertainties in the prediction of LIC from an RF classifier applied to optical remote sensing observations. RF is chosen not only due to its efficiency and predictive performance in the same classification problem as indicated by previous research [10], but also for its ability to measure and quantify predictive uncertainty. As will be explained in more detail later on, this is essentially due to the added information that is provided by an entire collection of predictions produced an ensemble technique such as RF, compared to just a single prediction obtained from a conventional classifier. Roughly speaking, the idea is that epistemic uncertainty should be reflected by the agreement or disagreement of the ensemble members: If all of them agree on more or less the same prediction, then this is a sign of low epistemic uncertainty. On the other side, a strong disagreement between the predictions can be taken as an indicator of high epistemic uncertainty. Similarly, information about aleatoric uncertainty can be extracted from the collection of predictions. This article describes the methodology for calculating uncertainties and outlines a strategy for generating lake ice maps for each observation. Using optical sensors with high revisit rates provides observations at a location with overlapping swaths. The lake ice map product can result from the fusion of multiple maps and the fusion can be done at the data or model level with aleatoric or epistemic uncertainties as the main criteria.

The rest of this article is organized as follows. Section I describes the introduction. Section II begins with an overview of the study area and datasets employed in our research. Section III introduce the methodology for the classification of pixels and the quantification of both aleatoric and epistemic uncertainties for such classifications. Sections IV and V analyze the results at both the pixel and window patch levels, followed by rejection criteria analysis. Finally, Section VI concludes this article.

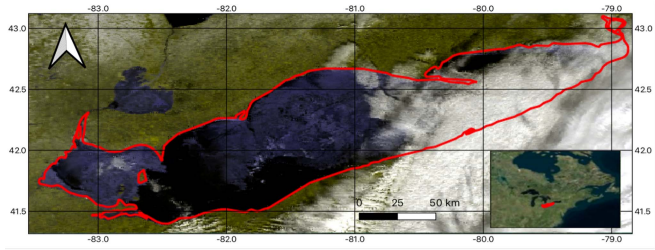

Fig. 1. False color RGB composite (R: band 2, G: band 2, B: band 1) of MODIS Terra on 2014/01/14 covering our study area, Lake Erie.

II. Data and Study Area

MODIS Terra Level 1B calibrated radiances product (MOD02), Collection 6.1 (top-of-atmosphere reflectance) was utilized for mapping lake ice, water, and cloud and uncertainty. MOD02QKM bands 1-2 with a 250 m pixel spacing and MOD02HKM bands 3-7 with a 500 m pixel spacing were used. We applied methods proposed by [15] and [10] for resampling to 250 m pixels and for the optimal combination of bands, respectively.

We selected three winters (2014, 2016, and 2018) from Lake Erie, one of the Laurentian Great Lakes of North America (see Fig. 1), for analysis. Lake Erie covers an area of 25 655 km^2, with an average depth of 19 m. Lake Erie is an exception to the other four Great Lakes (Huron, Michigan, Ontario, Superior) as it sometimes completely freezes over during winter due to its shallow depth. Winter 2014 is one of those instances whereby the Great Lakes experienced their second-largest ice coverage since 1973, due to persistent low air temperatures (NOAA/GLERL NOAA Great Lakes Environmental Research Laboratory—Historical Ice Cover. Available online: https://www.glerl.noaa.gov/data/ice/historical). MODIS images were filtered to select the ones with less than 50% cloud coverage for further uncertainty analysis and to minimize the imbalanced impact of the number of pixels in each class. From January to April of 2014,[1] 2016,[2] and 2018[3], 12, 10, and 19 images were selected, respectively. To train, validate, and test the RF model, extensive data labels are collected to maintain spatial and temporal independence. Additional independent labels collected in the study area are for uncertainty rejection analysis. The information on data labels is provided in Section IV-A.

[1]2014/01/03, 2014/01/09, 2014/01/14, 2014/01/29, 2014/02/12, 2014/03/07, 2014/03/13, 2014/03/30, 2014/03/18, 2014/04/06, 2014/04/12, and 2014/04/17.

[2]2016/01/19, 2016/01/24, 2016/01/28, 2016/02/05, 2016/03/17, 2016/04/27, 2016/01/24, 2016/02/01, 2016/02/06, and 2016/04/17.

[3]2018/01/04, 2018/01/13, 2018/01/28, 2018/02/26, 2018/03/19, 2018/01/05, 2018/01/19, 2018/02/12, 2018/03/02, 2018/03/25, 2018/01/06, 2018/01/20, 2018/02/13, 2018/03/15, 2018/04/23, 2018/01/09, 2018/01/26, 2018/02/14, and 2018/03/18.

III. Methodology

In this section, we briefly recall the ML and uncertainty quantification methodology we build on. We describe the formalization of the predictive modeling task as a problem of supervised learning, its instantiation with RFs as an ensemble-based learning algorithm, and the quantification of uncertainty in terms of appropriate numerical measures. Our approach is largely based on the work by [16].

A. ML for Predictive Modeling

Consider a standard supervised learning (classification) setting, in which a learner is given access to a set of training data

$$\mathcal{D}\,\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N} \subset \mathcal{X} \times \mathcal{Y} \tag{1}$$

where $\mathcal{X}$ is an instance space and $\mathcal{Y}$ a finite set of K possible class labels that can be associated with an instance. In our study, the set of class labels is given by

$$\mathcal{Y} = \{y_{\text{water}}, y_{\text{ice}}, y_{\text{cloud}}\} \tag{2}$$

and each instance $\boldsymbol{x} \in$ is a feature vector representing a pixel (cf. Section II). Assuming that the training examples are generated (independently) according to an underlying joint probability distribution P on $\mathcal{X} \times \mathcal{Y}$, i.e., that each $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ is a realization of $(X, Y) \sim P$, the task of predictive modeling is to learn the marginal $P(Y \mid X)$. More specifically, we seek a mapping

$$h : \mathcal{X} \longrightarrow \mathbb{P}(\mathcal{Y}) \tag{3}$$

that associates with each $\boldsymbol{x} \in \mathcal{X}$ a predicted probability $\hat{P}(Y \mid X = \boldsymbol{x})$, which is an accurate approximation of the true marginal — the quality of the approximation is specified in terms of a so-called loss function. If $\mathcal{Y}$ is finite, then $\hat{P}(Y \mid X = \boldsymbol{x})$ can be represented as a probability vector $h(\boldsymbol{x}) = (p_1, \ldots, p_K)$, where p_k is the probability predicted by h for the kth class label.

In addition to the training data $\mathcal{D}$, the learning algorithm is given a *hypothesis space* $\mathcal{H}$, from which a predictor (hypothesis) h can be chosen. In our case, $\mathcal{H}$ consists of all predictors $\mathcal{X} \longrightarrow \mathbb{P}(\mathcal{Y})$ that can be represented by a decision tree. The learner's selection of a specific predictor h is guided by an appropriate induction principle. The construction of decision trees, for example, is guided by the idea of uncertainty (entropy) reduction.

B. Uncertainty Quantification

Suppose a predictor h has been trained. How can we quantify the uncertainty of a prediction $h(\boldsymbol{x})$ for a specific instance $\boldsymbol{x}$? The goal of uncertainty quantification is to provide numerical measures of the overall (total) uncertainty as well as the aleatoric and epistemic uncertainty of the prediction.

As a prediction $h(\boldsymbol{x})$ in the form of a probability distribution on $\mathcal{Y}$ can only capture aleatoric uncertainty, the representation of epistemic uncertainty requires the learner to go beyond the induction of a single predictor h. In one way or the other, it needs to represent its uncertainty about (the predictions produced by) h. In the Bayesian approach to ML, this is accomplished by a (second-order) probability distribution Q on the hypothesis space $\mathcal{H}$. Thus, instead of inducing a single predictor, a Bayesian learner maintains a probability distribution over the entire hypothesis space, and learning essentially consists of replacing a prior on this space by a posterior distribution. The more concentrated this distribution becomes, the less (epistemically) uncertain the learner is.

On the basis of the learner's "belief" Q, uncertainty measures can be derived in various ways. The most common approach in ML relies on Shannon entropy as an established measure of (total) uncertainty, and leverages the information-theoretic result that entropy can be expressed as the sum of conditional entropy and mutual information [17]. Thus, the total uncertainty of the prediction $h(\boldsymbol{x})$ is the entropy

$$\text{TU}(\boldsymbol{x}) = H[\,\bar{\boldsymbol{p}}\,] = -\sum_{k=1}^{K} \bar{p}_k \cdot \log_2 \bar{p}_k \tag{4}$$

where H denotes Shannon entropy and the posterior predictive distribution $\bar{\boldsymbol{p}} = (\bar{p}_1, \ldots, \bar{p}_K)$ is obtained through (Bayesian) model averaging, i.e., averaging the probability predictions $\boldsymbol{p} = h(\boldsymbol{x})$ made by the individual models $h \in \mathcal{H}$, weighted by their posterior probability

$$\bar{\boldsymbol{p}} = \int_{\mathcal{H}} h(\boldsymbol{x})\, d\,Q(h)\,. \tag{5}$$

As fixing a single model h means committing to a single predictive distribution and hence removing all epistemic uncertainty, the entropy of such a distribution is a suitable measure of aleatoric uncertainty. The conditional entropy is the expectation of this measure with regard to Q

$$\text{AU}(\boldsymbol{x}) = \int_{\mathcal{H}} H\,[h(\boldsymbol{x})]\; d\,Q(h)\,. \tag{6}$$

Epistemic uncertainty can then be computed as the difference $\text{EU}(\boldsymbol{x}) = \text{TU}(\boldsymbol{x}) - \text{AU}(\boldsymbol{x})$. Thus defined, it coincides with the mutual information of the predictor h and outcome y (both considered as random variables). Intuitively, this appears to be plausible: epistemic uncertainty represents the (expected) reduction of uncertainty that is achieved by revealing the (uncertain) predictor h [12].

For complex model classes, the above approach is computationally intractable, due to the need for integrating over $\mathcal{H}$. In practice, ensemble methods are commonly used as an approximation [18]. Such methods train a set of M different predictors $h_1, \ldots, h_M \in \mathcal{H}$, which are considered as a representative sample from the true distribution Q. In our case, we instantiate this approach with decision trees as the model class, i.e., each ensemble members h_m is a decision tree.

Given a trained ensemble, the posterior predictive distribution as well as the uncertainty measures can be obtained through arithmetic averaging instead of integration: For a query instance $\boldsymbol{x}$, let $\boldsymbol{p}_m = (p_{1,m}, \ldots, p_{K,m})$ denote the probability distribution predicted for $\boldsymbol{x}$ by the mth ensemble member h_m. The posterior predictive distribution is then given by $\bar{\boldsymbol{p}} = (\bar{p}_1, \ldots, \bar{p}_K)$,

where

$$\bar{p}_k = \frac{1}{M} \sum_{m=1}^{M} p_{k,m} \,. \tag{7}$$

Moreover, total, aleatoric, and epistemic uncertainty are given as follows:

$$\mathrm{TU}(\boldsymbol{x}) = H\left[\bar{\boldsymbol{p}}\right] \tag{8}$$

$$\mathrm{AU}(\boldsymbol{x}) = \frac{1}{M} \sum_{m=1}^{M} H\left[\boldsymbol{p}_m\right] \tag{9}$$

$$\mathrm{EU}(\boldsymbol{x}) = \mathrm{TU}(\boldsymbol{x}) - \mathrm{AU}(\boldsymbol{x}). \tag{10}$$

C. Implementation With RFs

The approach outlined above has been realized for the RF classifier [19] as an ensemble method and implemented in Python. An RF is comprised of a collection of decision trees [20]. The primary objective of a decision tree is to predict the target variable by learning decision rules derived from the data features. In this context, each inner node within a tree is associated with a specific attribute or feature of the training data. The branches emanating from the node delineate potential outcomes or values of that feature, guiding the traversal to subsequent nodes until a final decision is ascertained at the terminal nodes, commonly referred to as leave nodes. In the classification setting, to derive the predicted probability distribution, one can utilize the relative frequency of the samples of each class within the leaf node.

To introduce diversity in the ensemble, RF trains each decision tree on a bootstrap, which involves randomly selecting data points from the original dataset with replacement [21]. Additional variability is injected into the training of decision trees by randomly choosing a subset of features as candidates for defining a split at an inner node of the tree, thereby ensuring greater diversity among the trees. Finally, the output of an RF is an aggregation of the outputs from the set of grown trees. In the classification setting, this aggregation typically involves computing the average (1) over the probability distributions from the individual trees, and this is the approach adopted in the present study.

IV. Results

A. Experimental Protocol

RF hyperparameters to be defined are the number of variables and the number of trees that are set to develop independent classifiers. In our RF classifier, MODIS TOA reflectance bands correspond to the predictor variables. The range of suitable hyperparameter values is provided by [10] based on analyzing gained accuracy. The RF model was trained and tested on an extensive, yet independent dataset collected from MODIS on several lakes worldwide above 40° latitude, during the years 2010–2020 with 1 048 575 labels. This independent dataset ensures that the model is generalized and not biased toward specific test lakes, such as Lake Erie. Before fine-tuning the hyperparameters, 20% of the data, equivalent to 209 715 data points, was set aside. Then, a 70%–30% train-validation split was applied to build the RF classifier using the aforementioned parameters. The hyperparameter values were sampled and tested to find the optimal settings, resulting in maximum classification accuracy. Ultimately, the hyperparameters were set to a maximum depth of 30, a maximum of three features, and 1000 estimators. Fig. 2 shows the data processing chain to map lake ice and uncertainties using TOA observation from MODIS sensor. We only used scenes with lake coverage to avoid mosaicing and to simplify the preprocessing steps.

To evaluate uncertainty, we created a new set of labels in Lake Erie and used this dataset for accuracy rejection analysis. Separating the datasets for this experiment helps ensure robustness by evaluating the model on new data, providing a more accurate assessment of its generalization performance.

B. RF Maps and Uncertainty Analysis

The RF model achieved a high accuracy score of 97% on the test set, which comprised 209 715 labels that were kept unseen during the hyperparameter fine-tuning step. The F-scores for the classification of ice, water, and cloud classes were 0.94, 0.97, and 0.98, respectively. Total, aleatoric, and epistemic uncertainty for each pixel was calculated using the methodology outlined in Section III. These uncertainties were then mapped in the same coordinate system as MODIS observations for further visual and statistical analysis. An example of the produced LIC map and corresponding uncertainties on 2014/01/03 are presented in Fig. 3. Noteworthy, as the uncertainty is calculated using entropy, it translates to the range of uncertainties as the minimum and maximum of entropy for probability distributions. The range of the uncertainty in RF is then defined as [0, log2(K)] with K being the number of classes. Therefore, in our experiments, the maximum uncertainty is $\log2(3) = 1.58$. The mean and variance of values of uncertainties for each year are reported in Table I. During the years of study, average cloud coverage in selected scenes was at a minimum in the winter of 2014. This year was marked as a unique winter season when the lake experienced complete ice coverage, as confirmed by other remotely sensed observations such as SAR [22], [23], it had the least average ice coverage compared to other selected years.

Visual inspections of satellite imagery across various dates reveal how the physical properties of lakes influence pixel classification. The observed differences in coverage indicate that assessments based on imbalanced class coverage are unreliable. For instance, the year with the lowest coverage exhibited the highest total uncertainty in RF classification. Noticeable examples are cloud in 2014 and water classes in 2016, which show the highest uncertainties while having the least coverage among other classes of the same winter season. On the other hand, the average epistemic uncertainties are lower than aleatoric uncertainties, so total uncertainty is mainly affected by aleatoric uncertainties. Interestingly, visual inspections of over 30 days show that patterns of both aleatoric and epistemic uncertainties are similar within a small-scale area of the lake's coverage, as can be seen in Fig. 3. The discrepancies between aleatoric and epistemic uncertainties become noticeable at smaller scales as explained in the next Subsection.

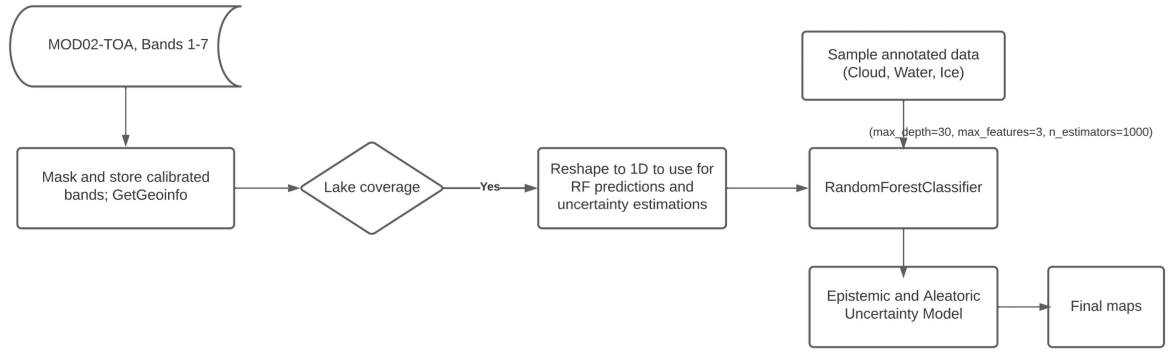

Fig. 2. General flowchart of data processing steps from extracting TOA bands of MODIS based on lake coverage to RF and uncertainty maps.

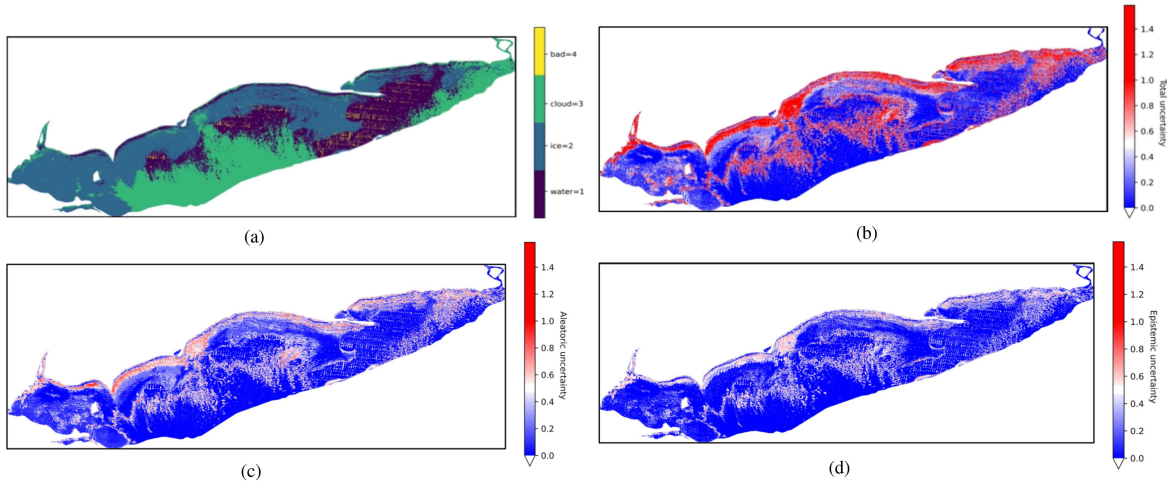

Fig. 3. (a) RF LIC map of 2014/01/03. (b) Total uncertainty. (c) Aleatoric uncertainty. (d) Epistemic uncertainty.

TABLE I
AVERAGE UNCERTAINTIES CALCULATED FOR EACH CLASSIFIED PIXEL DURING WINTERS OF 2014, 2016, AND 2018 (AVERAGE IS APPLIED TO ALL PIXELS FOR EACH CLASS IN ALL SELECTED SCENES OF EACH YEAR)

		Coverage %	Total Uncertainty		Epistemic Uncertainty		Aleatoric Uncertainty	
			Mean	Variance	Mean	Variance	Mean	Variance
2014	Water	61.7	0.38	0.38	0.16	0.19	0.22	0.21
	Ice	24.9	0.36	0.38	0.17	0.21	0.18	0.19
	Cloud	12.4	0.61	0.49	0.27	0.24	0.33	0.26
		Coverage %	Total Uncertainty		Epistemic Uncertainty		Aleatoric Uncertainty	
			Mean	Variance	Mean		Mean	Variance
2016	Water	0.5	0.92	0.42	0.42	0.23	0.5	0.21
	Ice	62.6	0.24	0.34	0.11	0.17	0.13	0.17
	Cloud	33.4	0.28	0.41	0.13	0.21	0.15	0.21
		Coverage %	Total Uncertainty		Epistemic Uncertainty		Aleatoric Uncertainty	
			Mean	Variance	Mean	Variance	Mean	Variance
2018	Water	42.6	0.30	0.38	0.17	0.18	0.18	0.20
	Ice	35.7	0.37	0.42	0.17	0.21	0.20	0.22
	Cloud	20.2	0.39	0.48	0.12	0.22	0.22	0.26

C. Neighborhood Analysis

Consistent variances of epistemic uncertainties tell us that the classification model is not out-performing or miss-performing in classifying any specific class. In general, these average uncertainty values cannot infer specificity on class performance as differences in values are not significant. Furthermore, uncertainties in each pixel reported in Table I do not include any spatial context. To take spatial variability into account, we studied window-based statistics with a 5 by 5 moving window to analyze how variability in surrounding pixels impacts uncertainties. The histograms of each class's pixel counts based on grouping into intervals of 10 were analyzed and it was observed that employing a natural breaks classification with intervals of 10 and 20 effectively reduces variance within classes and maximizes variance between classes. Consequently, for the window-based statistical analysis of uncertainties within a 5×5 window, three groups of pixel counts—namely 10, 10–20, and 20–25—have been selected.

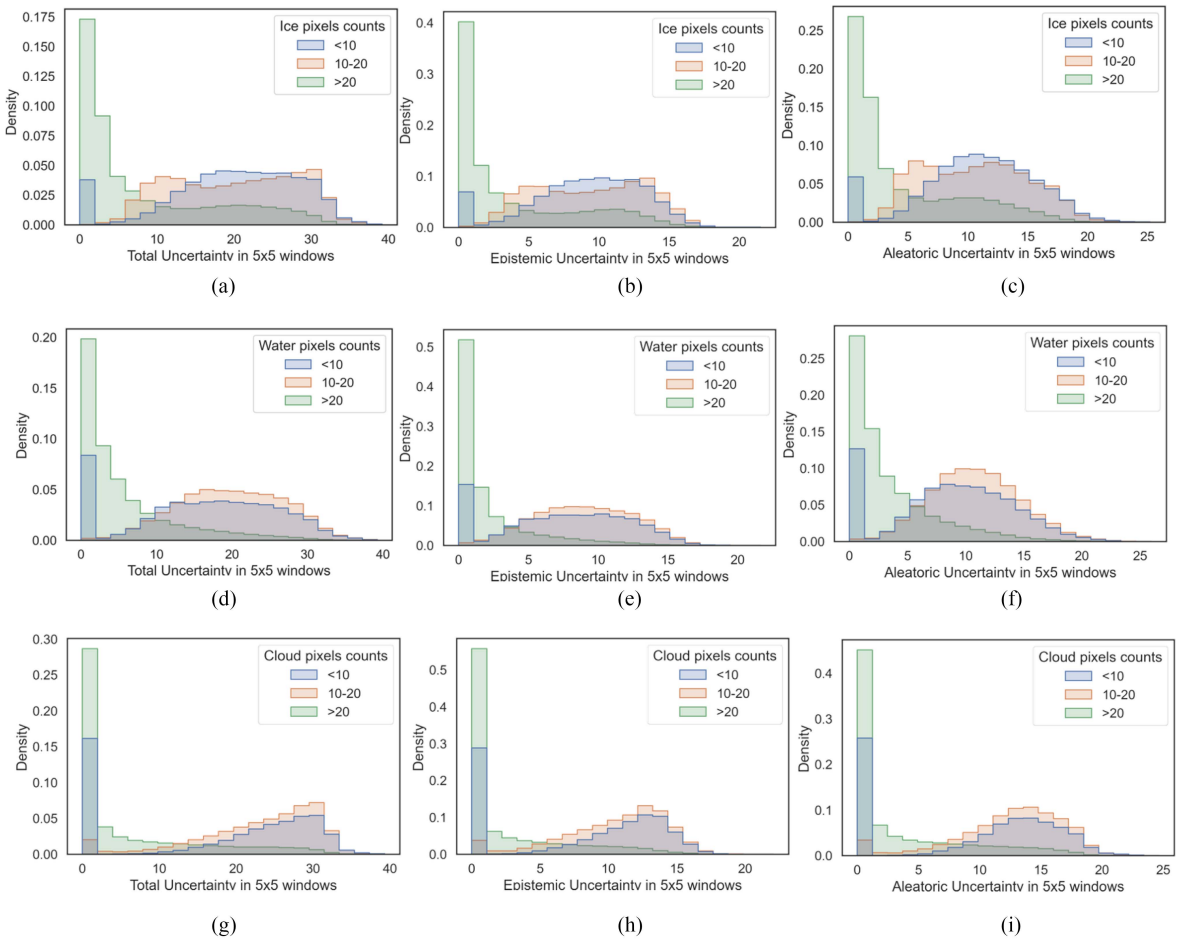

Fig. 4. Density histograms of uncertainty are plotted for selected scenes, categorizing each RF class in the center of a 5 by 5 window based on the total number of instances of that class within the window. (a), (b), and (c) show total, epistemic, and aleatoric uncertainties for the ice class, whereas (d), (c), and (f) maps out uncertainties for the water class, and (g), (h), and (i) represent uncertainties for the cloud class in the same order. The peak for pixel counts less than 10 in the lowest uncertainty bin indicates relatively isolated pixels of a given class surrounded by a majority of another class.

Uncertainties were calculated for three groups (<10, 10–20, >20) of classified ice, water, and cloud pixels within 5 by 5 windows, and their density histograms are plotted in Fig. 4. As can be seen, the uncertainties of classified ice, water, and cloud decrease when there is more homogeneity of the same class within 5 by 5 windows. A similar pattern can be seen in all classes for aleatoric and epistemic uncertainties, and there is no obvious difference between uncertainties' dispersion with pixel counts with less than 10 or 10 to 20 counts. In other words, as long as there are less than 20 pixels of the same class within a 5 by 5 window, the uncertainties are normally distributed and the mean is much higher than in cases where more than 20 pixels of the same class are mapped. The high density of the first bin of histograms in all three classes for both aleatoric and the epistemic uncertainties with less than 10 counts of the class of interest, can be explained by the presence of regions, where a low number of counts of a specific class in the center exists and a large number of the same class of another class is present. For instance, in the case of plot Fig. 4(d), we have a large number of windows with a small fraction of water pixels surrounded by clouds or ice with more than 20 pixels and as a result, the first bin with the lowest total uncertainty values has a high frequency of occurrence. This highlights the importance of spatial variability, as both aleatoric and epistemic uncertainties are found to be higher in windows where a small pixel count of the class of interest is surrounded by other classes.

D. Accuracy Rejection Experiment

To assess the effectiveness of uncertainty quantification, accuracy–rejection curves are plotted. Here, we arrange our RF predictions by their uncertainty values and progressively discard the uncertain ones while monitoring the accuracy of the remaining predictions. The underlying concept is that if the uncertainty quantification reliably distinguishes highly uncertain predictions (likely incorrect) from highly certain ones (mostly correct), the model's accuracy will improve as more uncertain predictions are rejected. Conversely, if predictions are rejected randomly, there will be no impact on accuracy.

Accuracy–rejection curves for a specific scene with balanced coverage of ice and water (with cloud coverage of only 0.35% in the scene) are shown in Fig. 5. These plots depict the F1-accuracy of classifying ice and water, with respect to rejection, on the

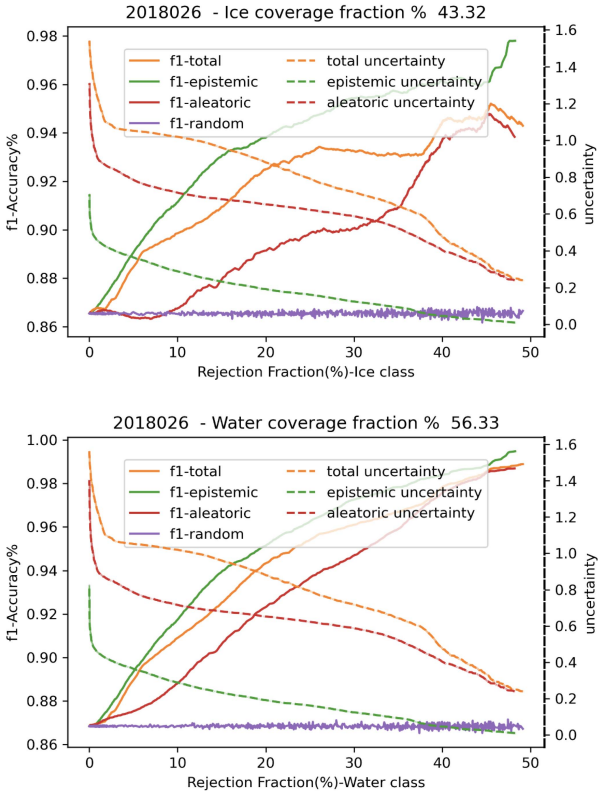

Fig. 5. Rejection analysis based on aleatoric, epistemic, and total uncertainties for ice and water classification.

left *y*-axis, correlated with rejected pixels exhibiting higher uncertainties, encompassing total, epistemic, and aleatoric uncertainties. The uncertainty values are represented by dashed lines in alignment with the right *y*-axis. As anticipated, there is an increase in accuracy performance for all three types of uncertainty rejections—total, epistemic, and aleatoric—while random rejection stays unchanged. According to these plots, epistemic uncertainty demonstrates the most effective rejection performance. This suggests that by rejecting fewer predictions, we can achieve the same level of accuracy as with total or aleatoric uncertainty, which will make more rejections. This discovery also validates the utility of decomposing total uncertainty into epistemic and aleatoric components when determining LIC maps.

V. Discussion

Visual inspections of the classified maps and corresponding uncertainty maps reveal that the highest uncertainties are present along edges/transitions between water, ice, and cloud cover. In addition, based on our prior knowledge of Lake Erie's ice coverage [23], high uncertainty was also found in areas with thin ice cover, which are present in the western basin of Lake Erie. These findings can help feed more informative annotations for RF training to gain more accurate classification maps. Misidentification of surface types under conditions of variable cloud cover was another noteworthy observation. These cases frequently result in the incorrect classification of other surface features. An example of this is observed on January 21, when ice was mistakenly identified as cloud cover. The uncertainty values for this date were markedly higher compared to those of adjacent dates with correct classifications, underscoring the utility of uncertainty analysis. Another significant application would be using uncertainty maps as label uncertainties while training an ML-based model. The outcome of such an application would contributes to mapping at scale with minimal resources for annotations, similar to weakly supervised approaches.

To better understand the spatial pattern of the uncertainties, rather than only inferring statistics based on a pixel-based averaging in the mapped area, we used window-based statistics to estimate uncertainties of classes in each window while considering class variabilities in each window. Results indicate that spatial variability is one of the main drivers of higher uncertainty in both epistemic and aleatoric forms. This work can be applied in the collective classification technique [24], where the pixel to be classified is compared with its neighbors iteratively. By incorporating uncertainties in collective classification and taking into account the spatial context within a region, the accuracy of classification could be improved.

This investigation represents a significant shift from previous methodologies in the domain of lake and sea ice mapping, which predominantly utilized convolutional neural networks (CNNs) and employed approaches such as modified loss functions [23], [25], uncertainty decomposition [13], or calibrated probabilities [26]. In contrast, the use of RF in this study for analyzing multispectral data introduces a methodologically rigorous and principled approach that is adaptable to other research scenarios involving similar data types. This strategy enhances the robustness of uncertainty quantification, facilitating more reliable classification outcomes and offering potential frameworks for future applications.

VI. Conclusion

Mapping of LIC using optical spaceborne sensors has a long history, however, quantification of how reliable are those retrievals at the pixel level is lacking. Uncertainty estimates expand the lake ice map product usability by making researchers aware of aleatoric and epistemic uncertainty for incorporating ice fractions in numerical models, such as lake and weather forecasting models. Quantifying both aleatoric and epistemic uncertainties are crucial for improving the reliability and robustness of model predictions. Aleatoric uncertainties, which represent the inherent variability and randomness within the data or system being modeled, remain irreducible regardless of the amount of information available. Epistemic uncertainties are essential for comprehending a model's performance and are integral to the fusion of disparate models for generating a specific product. Uncertainty incorporation can be done in the form of direct integration of observation error variance or as a quality control flag. To expand the application of the presented work, efforts are underway to integrate this methodology within the existing processing chain of daily LIC product generation of ESA's CCI Lakes project. This integration will include the use

of uncertainty as an informative tool to identify problematic classifications, both spatially and temporally, within the LIC product. Such integration will not only extend the current work, but also address data limitations, thereby providing a foundation for global lake ice mapping. The availability of uncertainty data will facilitate the development of data fusion methods, leveraging discrepancies within classifications or lakes and across different dates to enhance the overall product quality.

ACKNOWLEDGMENT

The authors would like to thank YuhaoWang for his help with data annotations for training the Random Forest classification model.

REFERENCES

[1] X. Yang, T. M. Pavelsky, L. P. Bendezu, and S. Zhang, "Simple method to extract lake ice condition from landsat images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4202010. [Online]. Available: https://doi.org/10.1109/tgrs.2021.3088144

[2] W. Han, C. Huang, J. Gu, J. Hou, and Y. Zhang, "Spatial-temporal distribution of the freeze-thaw cycle of the largest lake (Qinghai lake) in China based on machine learning and MODIS from 2000 to 2020," *Remote Sens.*, vol. 13, no. 9, 2021, Art. no. 1695.

[3] L. C. Brown and C. R. Duguay, "Lake ice," 2022. [Online]. Available: https://repository.library.noaa.gov/view/noaa/48536

[4] X. Wang et al., "High-resolution mapping of ice cover changes in over 33,000 lakes across the north temperate zone," *Geophys. Res. Lett.*, vol. 48, no. 18, 2021, Art. no. e2021GL095614.

[5] S. Zhang, T. M. Pavelsky, C. D. Arp, and X. Yang, "Remote sensing of lake ice phenology in Alaska," *Environ. Res. Lett.*, vol. 16, no. 6, 2021, Art. no. 064007.

[6] J. Kropáček, F. Maussion, F. Chen, S. Hoerz, and V. Hochschild, "Analysis of ice phenology of lakes on the Tibetan plateau from modis data," *Cryosphere*, vol. 7, no. 1, pp. 287–301, 2013.

[7] Y. Cai, C.-Q. Ke, X. Li, G. Zhang, Z. Duan, and H. Lee, "Variations of lake ice phenology on the Tibetan plateau from 2001 to 2017 based on modis data," *J. Geophysical Res.: Atmospheres*, vol. 124, no. 2, pp. 825–843, 2019.

[8] J. Murfitt, C. R. Duguay, G. Picard, and G. E. Gunn, "Investigating the effect of lake ice properties on multifrequency backscatter using the snow microwave radiative transfer model," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4305623.

[9] K. Barbieux, A. Charitsi, and B. Merminod, "Icy lakes extraction and water-ice classification using landsat 8 OLI multispectral data," *Int. J. Remote Sens.*, vol. 39, no. 11, pp. 3646–3678, 2018.

[10] Y. Wu, C. R. Duguay, and L. Xu, "Assessment of machine learning classifiers for global lake ice cover mapping from MODIS TOA reflectance data," *Remote Sens. Environ.*, vol. 253, 2021, Art. no. 112206.

[11] L. Carrea et al., "Satellite-derived multivariate world-wide lake physical variable timeseries for climate studies," *Sci. Data*, vol. 10, no. 1, p. 30, 2023.

[12] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods," *Mach. Learn*, vol. 110, no. 3, pp. 457–506, 2021.

[13] X. Chen, K. A. Scott, L. Xu, M. Jiang, Y. Fang, and D. A. Clausi, "Uncertainty-incorporated ice and open water detection on dual-polarized SAR sea ice imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5201213.

[14] K. A. Scott, L. Xu, and H. K. Pour, "Retrieval of ice/water observations from synthetic aperture radar imagery for use in lake ice data assimilation," *J. Great Lakes Res.*, vol. 46, no. 6, pp. 1521–1532, 2020.

[15] A. P. Trishchenko, Y. Luo, and K. V. Khlopenkov, "A method for downscaling MODIS land channels to 250-m spatial resolution using adaptive regression and normalization," *Proc. SPIE*, vol. 6366, pp. 46–53, 2006.

[16] M. Shaker and E. Hüllermeier, "Aleatoric and epistemic uncertainty with random forests," in *Proc. IDA, 18th Int. Symp. Intell. Data Anal. ser. LNCS*, Konstanz, Germany: Springer, 2020, vol. 12080, pp. 444–456.

[17] S. Depeweg, J. Hernandez-Lobato, F. Doshi-Velez, and S. Udluft, "Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning," in *Proc. Int. Conf. Mach. Learn.*, Stockholm, Sweden, 2018.

[18] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[19] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[20] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, pp. 81–106, 1986.

[21] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, pp. 123–140, 1996.

[22] J. Wang, C. R. Duguay, D. A. Clausi, V. Pinard, and S. E. Howell, "Semi-automated classification of lake ice cover using dual polarization RADARSAT-2 imagery," *Remote Sens.*, vol. 10, no. 11, 2018, Art. no. 1727.

[23] N. Saberi, K. A. Scott, and C. Duguay, "Incorporating aleatoric uncertainties in lake ice mapping using RADARSAT–2 SAR images and CNNs," *Remote Sens.*, vol. 14, no. 3, 2022, Art. no. 644.

[24] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad, "Collective classification in network data," *AI Mag.*, vol. 29, no. 3, pp. 93–93, 2008.

[25] X. Chen, R. Valencia, A. Soleymani, and K. A. Scott, "Predicting sea ice concentration with uncertainty quantification using passive microwave and reanalysis data: A case study in Baffin bay," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4300213.

[26] T. Wulf, J. Buus-Hinkler, S. Singha, H. Shi, and M. B. Kreiner, "Pan-arctic sea ice concentration from sar and passive microwave," *EGUsphere*, vol. 2024, pp. 1–34, 2024.

Nastaran Saberi (Member, IEEE) received the B.S. and M.S. degrees in geomatics engineering from the University of Tehran, Tehran, Iran, in 2008 and 2011, respectively, and the Ph.D. degree in geography from the University of Waterloo, Waterloo, ON, Canada, in 2019.

She previously held academic and research positions with the University of Waterloo and is now the Co-Founder and CTO of GeoMate, where she leads the development of AI-driven mapping technologies for smart mobility and autonomous systems. Her current work focuses on deep learning applied to geospatial data and models, with an emphasis on enhancing the analysis and utility of satellite-based data for smart mobility applications.

Mohammad Hossein Shaker received the B.Sc. degree in computer science from the University of Qom, Qom, Iran, and the M.Sc. degree in computer science from Isfahan University of Technology, Isfahan, Iran. He is currently working toward the Ph.D. degree with the Chair of Artificial Intelligence and Machine Learning (AIML), Ludwig Maximilian University of Munich (LMU), Munich, Germany, under the supervision of Prof. Eyke Hüllermeier.

His research interests include uncertainty quantification and calibration in machine learning and artificial intelligence.

Claude R. Duguay (Member, IEEE) received the B.Sc. degree from the Université de Montréal, Montreal, QC, Canada, in 1983, the M.A. degree from The State University of New York at Buffalo, Buffalo, NY, USA, in 1985, and the Ph.D. degree from the University of Waterloo, Waterloo, ON, Canada, in 1989.

He is currently a Professor with the Department of Geography and Environmental Management, University of Waterloo. He is also the Founder and Scientific Advisor of H2O Geomatics Inc., a research spin-off from the University of Waterloo. His current research focuses on the development of physically-based and machine-learning algorithms and lake products from microwave and optical satellite data and the improvement of the representation of cryospheric processes in lake model schemes implemented in numerical weather prediction and climate models. His research interests include remote sensing, field observations, and modeling of freshwater ice to deepen our understanding and predictive capabilities of lake atmosphere interactions in cold climate regions.

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, Canada, in 2001, and the Ph.D. degree from the University of Waterloo, Waterloo, ON, Canada, in 2008, all in mechanical engineering.

In 2008, she was a Postdoctoral Researcher with the Data Assimilation and Satellite Meteorology Research Section, Environment and Climate Change Canada, Toronto, ON, Canada, 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, moving to the Department of Mechanical and Mechatronics Engineering in 2023. She works on remote sensing, deep learning and computational fluids.

Eyke Hüllermeier (Senior Member, IEEE) received the Ph.D. degree in computer science from Paderborn University, Paderborn, Germany, in 1997, and a Habilitation degree in 2002.

He is currently a Full Professor with the Institute of Informatics, University of Munich, Germany, where he heads the Chair of Artificial Intelligence and Machine Learning. Prior to his current position, he held professorships with several other German universities and was a Marie Curie fellow with the Institut de Recherche en Informatique de Toulouse (France). His research interests are centered around methods and theoretical foundations of artificial intelligence, with a specific focus on machine learning and reasoning under uncertainty.

Dr. Hüllermeier is President of the European Association for Data Science, an Editor-in-Chief of Data Mining and Knowledge Discovery, and serves on the editorial board of various journals, including IEEE TRANSACTION ON PATTERN RECOGNITION AND MACHINE INTELLIGENCE.