# Assessment of categorical triple collocation for sea ice/open water observations: Application to the Gulf of Saint Lawrence

K. Andrea Scott *Member IEEE*

***Abstract*—Monitoring the sea ice cover is important for both climate studies and ice operations, such as shipping. It is challenging to validate even basic essential variables, such as the sea ice extent, due to a lack of appropriate validation data. Instead of focusing on validation, this study looks at the use of categorical triple collocation (CTC) for the task of quantitatively comparing three colocated data sets. CTC has been developed and used in earlier studies to rank binary data sets. Here we extend earlier studies, and bring in recent results from the binary classification community, to estimate the class imbalance (the relative proportion of each class, ice or water). We then use this class imbalance to obtain quantitative estimates of the proportion correct of ice (sensitivity) and the proportion correct of water (specificity). The methodology is first tested using toy data, after which three data sets from the Gulf of Saint Lawrence, on the east coast of Canada, are used. These data sets are from an ice-ocean model, a passive microwave sea ice concentration retrieval, and a sea ice concentration retrieval from SAR. By looking at both the sensitivity and specificity it is found the passive microwave data have difficulty recognizing ice during freeze-up, but perform well at obtaining the correct water observations. This distinction cannot be made by ranking the datasets. The CTC method is compared with, and found to be complementary to, a validation using ice/water states from the Interactive Multisensor Snow and Ice Mapping System.**

***Index Terms*—sea ice validation, triple collocation, binary data**

## I. Introduction

Binary data, or data that can have two states, arises frequently in geophysical applications. Some examples include, the freeze/thaw state of soil [21] and binary snow cover maps [27], both of which are important for agricultural applications; ice on/off dates that are used for monitoring changes in rivers and lakes [23], segmentation of a satellite image into a flooded vs non-flooded region [9], and classification of sea ice satellite imagery into ice and water [18, 17, 14]. It is this last application that is of interest here.

For sea ice monitoring, identification of ice and water in a given region is important because it is used to estimate the sea ice extent, and ice concentration. The sea ice extent (SIE), commonly defined as the area of the ocean covered by sea ice of concentratration 15% or greater, is monitored each year as a climate change indicator [8]. Annual minimum and maximum values of the SIE, and the spatial distribution of ice cover, are linked to ocean and atmospheric conditions, to provide an understanding of factors driving sea ice change in the Arctic. While the relevance of SIE in the Arctic is well known in the scientific community, ice/water observations at a smaller scale are equally important. Identification of openings in the ice cover is important for shipping, since ship captains prefer to travel in open water when possible. Such openings in the ice cover are also known to transfer large quantities of heat from the ocean to the atmosphere, and should be included in the bottom boundary condition used to provide numerical weather forecasts of Arctic regions. This is not current practice, as the ice conditions used in numerical weather prediction are typically from coarse-resolution (25 km) satellite data. As shown in [2], incorporating higher resolution information from satellite data leads to stronger and more localized fluxes. Further study and more realistic turbulence closures are needed to evaluate and understand the impact of these fluxes.

Different sources of satellite data of ice-covered regions can be used, either individually or together, to identify regions covered by ice or water. The most intuitive data sets are those from visible/infrared sensors. In the visible bands ice is typically bright and water is typically dark and a binary ice/water state can be obtained by thresholding reflectance values [33]. It can be challenging to define a single threshold that works over a range of ice types, for example, thin ice is often dark and can be mistaken for open water [26]. While data from the visible bands cannot be used during the polar winter due to the lack of sunlight, it is at these times that thermal bands can be used [19]. A major challenge however with using visible/infrared data is that clouds have a spectral signature similar to ice, and must be removed from the image before classification. It can be challenging to identify clouds in an automated manner, and this can lead to errors in the retrieved ice/water observations [19].

Satellite imagery obtained in the microwave imaging bands, or approximately 0.3-40 GHz for active microwave imaging and 1-200 GHz for passive microwave imaging, is widely used for sea ice monitoring as it does not suffer from the mentioned shortcomings of visible/infrared data. In particular, sunlight is not required and the microwave bands at low frequencies (1-37 GHz) are relatively insensitive to cloud cover. For passive microwave imaging, the challenge is the the large instrument field of view required, eg. nominally 5 km to 55 km for the Advanced Microwave Scanning Radiometer-2 (AMSR2) sensor. Much higher spatial resolution (e.g. 50 m) can be obtained using synthetic aperture radar (SARs), with the corresponding challenge being interpretation of the imagery. This has been a topic of research for some time, and

K. A. Scott is with the Department of Systems Design Engineering at the University of Waterloo, Waterloo, Canada (email: ka3scott@uwaterloo.ca)

at present, there are several methods that have been proposed for the problem of identification of ice and water regions in SAR imagery [14, 18, 17].

In addition to these data sets, which are based on output from a single sensor, there are also products that combine information from a variety of sensors. These are commonly developed in operational ice centres, and are based on manual interpretation of satellite imagery carried out by trained analysts. A prominent example of there products is the interactive multisensor snow and ice mapping system (IMS) [24]. This product combines data from a variety of satellite sensors and in-situ sources to provide a manual interpretation for which each ocean pixel is labelled as ice, water or cloud. To delineate between ice and water the analyst uses a threshold of 40%, i.e. each pixel that appears to have ice concentration greater than 40% is labelled as ice.

A major problem in the sea ice forecasting community is verification of these binary ice/water data sets. Typically, a single dataset, such as IMS, is chosen as the benchmark to which the others are compared. Clearly, this is problematic as the chosen benchmark itself may suffer from deficiencies, or it may itself be based on one of the datasets to be verified, and thus not independent. For example, the most recent version of IMS utilizes visual/infrared (VIS/IR) data, passive microwave data and SAR data [24]. Thus, when using IMS as a verifying dataset for a product incorporating one of those data sources the results must be interpreted carefully. For example when assessing scores for a data assimilation system analysis with and without data from a specific sensor, there is a potential for the analysis to be in stronger agreement with IMS in regions where that sensor is a key component in the IMS analysis, which could bias the perceived impact of data from that sensor on the analysis.

It has been shown that binary data sets can be effectively compared, or ranked, using the triple collocation method for categorical data, or categorical triple collocation (CTC) [21]. This method is similar to the classic triple collocation method [31] in that three colocated data sets are required, and differs in that it does not use an affine error model and does not require that the errors of all three datasets are independent of each other, and are independent of the true state [34]. Instead, CTC employs the weaker assumption that the errors of the three datasets are conditionally independent, given the true state [21]. The CTC method has been demonstrated in several studies as effective in ranking soil freeze/thaw states [21, 20]. Here, the ranking is given by the relative proportion of correct classifications of a given dataset as compared to the other two, although the numerical values of the proportion of correct classifications remain unknown.

The present study builds on what has been done using CTC by linking this work to recent work in the area of ranking binary classifiers for the problem of unsupervised classification. More specifically, in Section II we build on the analysis in [25] and [16] and investigate a method that can be used to not only rank three binary datasets, but can also be used to obtain quantitative estimates of the proportion of correct classifications, and the proportion correct of each class. We show that this method can be used for the general case of non-stationary data, which is one for which the class imbalance (this is a measure of the relative frequency of each class) is changing as a function of either time or space. In Section III we test the method on a toy problem and in Section IV, testing is carried out using real data. The ranking and scores of CTC for the case using real data are also compared with those obtained using IMS as verification data. Discussion and concluding remarks are given in Sections V and VI respectively.

## II. Formulation of the problem

The notation used here is consistent with previous literature on ranking of categorical classifiers [21, 25]. It is assumed there is a binary true state, denoted as $T$, with values $T \in [-1, 1]$. Here we will consider these values to correspond to two classes, ice and water. However, the notation is sufficiently general to be applied to other types of binary variables, as has been done in [25, 21, 16]. We assume we have three binary classifiers that can assign ice and water labels, and denote the output of the $i'th$ binary classifier as $X_i$, where $i = 1, 2, 3$. Note $X_i \in [-1, 1]$.

Typically, in a binary classification problem, we are interested in the balanced accuracy, which is denoted here as $\pi_i$. This is the average of the sensitivity, $\psi_i$, which is the proportion of samples with $T = 1$ that are labelled correctly by the $i'th$ classifier, and the specificity, $\eta_i$, which is the proportion of samples with $T = -1$ that are labelled correctly by the $i'th$ classifier. The balanced accuracy for the $i'th$ classifier can be written as

$$\pi_i = \frac{\psi_i + \eta_i}{2}. \tag{1}$$

In the sea ice community, sensitivity and specificity are commonly referred to as the proportion correct ice and the proportion correct water respectively. However, the balanced accuracy is different than the total proportion correct, which is used in the sea ice community [6]. The total proportion correct is the total number of ice and water points that are correctly classified, divided by the total number of ice and water points in the verification data. If one class is dominant (e.g. ice) then the influence of the misclassifications of the other class (e.g. water) will not be noticed in the score [20].

It has been shown in [25] that the balanced accuracy can be related to the elements of the covariance matrix of the classifier outputs, $\mathbf{Q} = cov(X_i, X_j)$. More specifically,

$$Q_{ij} = (1 - b^2)(2\pi_i - 1)(2\pi_j - 1). \tag{2}$$

Here $b$ is the class imbalance, ie. $b = Pr[T = 1] - Pr[T = -1]$ [25], where $Pr$ denotes the probability. For a sea ice state with the same number of water points as ice points, $b = 0$, whereas a state with all ice points or all water points would correspond to $b = 1$ and $b = -1$ respectively. Equation (2) is the same as that used in [21] when the substitution $b = 2p-1$ [25] is made, where $p$ is $Pr[T = 1]$. Instead of working with the covariance matrix, it can be convenient to form a vector that directly relates the elements of $Q_{ij}$ to the balanced accuracies [21],

$$\boldsymbol{v} = \sqrt{1 - b^2}(2\boldsymbol{\pi} - 1), \tag{3}$$

where

$$\boldsymbol{v} = \begin{bmatrix} \sqrt{\dfrac{Q_{12}Q_{13}}{Q_{23}}} \\ \sqrt{\dfrac{Q_{12}Q_{23}}{Q_{13}}} \\ \sqrt{\dfrac{Q_{13}Q_{23}}{Q_{12}}} \end{bmatrix}. \tag{4}$$

Given that the elements of the covariance matrix are known from the data itself, then the balanced accuracy of each classifier $\pi_i$ can be determined from (3), *if* the class imbalance, $b$, is also known. However, in practice the class imbalance is not known, since it is a property of the true state. As has been pointed out in earlier studies [21], it is possible to rank the datasets using their $v_i$ values without knowing $b$ and this will correspond to their ranking in terms of $\pi_i$ because the $\sqrt{1-b^2}$ term is common for all $v_i$, $i = 1..3$ in equation (3). However, in some cases more information than a ranking of the datasets is desired. For example, for the ice/water classification problem, an opening in the ice cover (i.e. misclassified ice) would generate a large heat flux to the atmosphere if the ice/water state was used to initialize a coupled ice-atmosphere model, while spurious ice over open water (i.e. misclassified water) would be a less significant problem, as this ice would melt if the sea surface temperature was warm. To identify the performance of the classifier for these two situations, it is the sensitivity and specificity of the classifier that are required.

The sensitivity and specificity represent two unknown quantities, and two independent pieces of information are required to estimate them. The vector with elements corresponding to the mean of each classifier output, $\boldsymbol{\mu}$ and the $\boldsymbol{v}$ vector, both of which can be estimated from the data, can be used to obtain the following expressions for sensitivity and specificity [16]

$$\boldsymbol{\psi} = \frac{1}{2}\left(1 + \boldsymbol{\mu} + \boldsymbol{v}\sqrt{\frac{1-b}{1+b}}\right) \tag{5}$$

$$\boldsymbol{\eta} = \frac{1}{2}\left(1 - \boldsymbol{\mu} + \boldsymbol{v}\sqrt{\frac{1+b}{1-b}}\right). \tag{6}$$

In these equations the factors that lead to high values for sensitivity and specificity can be noted. For example, if a dataset has more ice and therefore a higher mean value, $\mu$, the sensitivity will increase, whereas if it has less ice and a lower value (more negative) for $\mu$, the specificity will increase. This is modified by the ranking of the dataset, which is captured by $v_i$, and the term containing the class imbalance. If $b \approx 1$ (more ice) the ranking plays little role in the sensitivity, and a stronger role in the specificity, and if $b \approx -1$ (more water) it plays more of a role in sensitivity and less in specificity.

To obtain the estimates of sensitivity and specificity from equations (5) and (6) we still have the problem that the class imbalance $b$ is not known *a priori*. To estimate this third quantity, an additional piece of information is needed. Here we follow the method in [16] that utilizes the three-dimensional covariance tensor constructed from the classifier outputs. Denoting this tensor as $\mathbf{T}$, it has elements

$$T_{ijk} = E[(X_i - \mu_i)(X_j - \mu_j)(X_k - \mu_k)] \tag{7}$$

where $i \neq j \neq k$, $E$ denotes the expectation, and $\mu_i$ denotes the mean of the output from the $i'th$ classifier. For the case with three classifiers, considered here, there is only one non-zero element in $\mathbf{T}$. In this case the relationship between $T_{ijk}$ and the class imbalance can be obtained as [16]

$$T_{ijk} = \alpha(b) v_i v_j v_k \tag{8}$$

where $\alpha(b) = -2b/\sqrt{1-b^2}$ or $b = -\alpha/\sqrt{4+\alpha^2}$. For the case when there are only three classifiers there is only one non-zero element in $\mathbf{T}$ and a single solution for $\alpha(b)$, $b$ can be found directly. In the more general case with more than three classifiers the value of $\alpha$ can be obtained using a minimization method [16]. The estimate of $b$ will be sensitive to $\boldsymbol{v}$ and $\mathbf{T}$, which may be noisy when the number of samples is limited and/or the quality of the data is poor, as may be the case for remote sensing applications. There will also be limits of applicability of this method due to the appearance of the $(1 \pm b^2)$ term in the denominator of equations (5) and (6), which will result in a singularity when $b = \pm 1$.

An additional challenge with applying this method to determine the class imbalance (and thereby the sensitivity, specificity and balanced accuracy for each classifier) is that, as pointed out in [21], for geophysical applications the true state is a function of either time (as in a time series at a fixed location) or space (as in the image data used in sea ice remote sensing). For both of these scenarios, the class imbalance is no longer stationary, but varies as a function of time or space. It was shown in [21] for the non-stationary case an expression similar to (3) can be used to rank the classifiers. with the class imbalance $b$ replaced by the expectation of this variable with respect to time (or space, if the variable is non-stationary in the spatial domain). In Appendix A we show, following a similar methodology as in [21] that an expression for the expectation of the class imbalance can be obtained using the three-dimensional covariance tensor of the datasets that is similar to that for the stationary case. To be more explicit, we show that for the binary case

$$T_{ijk} = \alpha(E[b]) v_i v_j v_k, \tag{9}$$

where $\alpha(E[b]) = -2E[b]/\sqrt{1-E[b]^2}$ and $E[b]$ is the expectation of the class imbalance. This expression for $\alpha$ can then be used to estimate $E[b]$ [16],

$$E[b] = -\alpha/\sqrt{4+\alpha^2}. \tag{10}$$

To summarize, for the case with three classifiers, the following steps can be used to obtain an estimate of the class imbalance ($b$), the sensitivity ($\psi$) and the specificity ($\eta$)

1) Estimate $\boldsymbol{\mu}$, $\boldsymbol{v}$ and $T_{ijk}$ from the classifier outputs.
2) Use equation (9) to obtain an estimate of $\alpha$ and (10) to obtain an estimate of $E[b]$.
3) Use equations (5) and (6) to estimate $\boldsymbol{\psi}$ and $\boldsymbol{\eta}$ using $\boldsymbol{v}$ and replacing $b$ with $E[b]$. This is valid because equations (5) and (6) are obtained using the relationship between $\boldsymbol{v}$ and $\boldsymbol{\pi}$ in equation (3), which is written in the non-stationary case with $b$ replaced by $E[b]$ (see [21]).

The method shown here is subject (as is the analysis in [16, 21]) to the assumption that all classifiers are conditionally

independent given the true state. For the application of interest here, this means the probability of two classifiers making an erroneous classification, given that the true state is ice, are independent events, and similarly given that the true state is water. For example, if one classifier is a binary ice/water state from SAR data, it may produce an erroneous ice observation due to wind-roughening of open water. However, if another classifier is a thresholded ice concentration from passive microwave data, where weather filters are used in the retrieval, it is not very likely that the passive microwave retrieval will contain the same error for the same reason. In general, one cannot be certain that the conditional independence assumption is met rigourously. We note this has been assumed true in previous studies [20, 21], and alternative methods that relax this assumption [15] will be addressed in a future study.

## III. Experiments with toy model

To investigate the proposed approach to estimate $b$, two toy model scenarios are considered. The first scenario is identical to the one in [21], which uses a true state with class imbalance varying between -1 and 1, but with $E[b]$=0. While in [21] it is the ranking of the classifiers that is estimated, here the ability of the method to retrieve $E[b]$, $\psi$ and $\eta$ was examined, in particular as a function of the number of samples available. The second toy model scenario is designed to investigate the accuracy with which $\psi, \eta$ and $E[b]$ can be retrieved for true states having various values of $E[b] \neq 0$. Again, it is the ability of the method to retrieve $E[b]$, $\psi$ and $\eta$ that is of interest.

### A. First Scenario: $E[b] = 0$

This scenario is identical to that in [21] and is described briefly for completeness. The probability of the true state exhibits a cosinusoidal variation with time $t$, over a 52 week period, varying from $p(t) = 1$ to $p(t) = 0$ and then back to $p(t)$=1. Given the following relationship between $p(t)$ and $b(t)$ [25]

$$b(t) = 2p(t) - 1 \tag{11}$$

this means the class imbalance varies between $b(t) = -1$ and $b(t) = 1$ with $E[b(t)] = 0$ over the 52 week period, because the average of a cosine over its period is zero. This could be considered data from a location that is ice-covered at the beginning of a time series, evolves to an open water state at the middle of the time series, and then returns to being ice-covered at the end of the time series. In the interest of building on previous work, we use data sets with the same characteristics as in [21]

$$\begin{aligned} \psi_1 = 0.8, \quad \eta_1 = 0.6 \\ \psi_2 = 0.9, \quad \eta_2 = 0.7 \\ \psi_3 = 0.98, \quad \eta_3 = 0.88. \end{aligned}$$

To carry out the experiments, data sets with these specifications are sampled from the true state. These data sets consist of binary values, and in the context of the methodology presented here, can be considered classifier outputs. The two and three dimensional covariances are calculated from these data sets and are used to estimate $\alpha$ and from this the class imbalance, following the method described in the Section II. This estimate of the class imbalance is referred to as $E[\hat{b}]$. This value of $E[\hat{b}]$ is then used to estimate the sensitivity and specificity according to equations (5) and (6). Experiments were carried out with different numbers of samples in the true state. For example, we looked at 500, 1000 and 2000 samples. If we consider these samples to be obtained as a time series at a fixed location over a single year, this would correspond to approximately 10, 20 and 40 samples each week. We also used bootstrapping, sampling 1000 replicates from our dataset with replacement to obtain confidence intervals for our estimates, as has been done in other triple collocation studies [21, 7]. We refer to mean values as the mean over these samples.

Result from this toy experiment are shown in Fig 1. It can be seen that the ability of the method to estimate the mean value of sensitivity and specificity improves, and the width of the confidence intervals decrease, as the number of samples increases. Using 1000 samples generally results in mean values for sensitivity and specificity that are within $5\%$ of the true values, and this is considered sufficient for our application. The values of the estimated class imbalance (not shown) were found to be within 1% of the true value, with lower errors corresponding to cases with larger numbers of samples. In all cases, the correct ranking of the classifiers was obtained, where the ranking was obtained by the mean value of the estimated $\boldsymbol{v}$ vector. According to equation (3) the dataset with the highest $v_i$ value has the highest $\pi_i$ and therefore is ranked first, and so on. Here, data set 3 is ranked first, followed by data set 2 and then data set 1, in agreement with the values prescribed for $\psi_i$ and $\eta_i$.

### B. Scenario 2: $E[b] \neq 0$

This scenario consists of a true state where the time-varying probability of the true state $p(t)$, has the property $E[p(t)] \neq 0$, and therefore the class imbalance also has the property $E[b(t)] \neq 0$. This could be thought of as data from a location where there is on average more (or less) ice as compared to water over the sample period. To carry out a set of experiments, each with a nonzero $E[b(t)]$, for each experiment a set of samples is drawn from the true state, for which $b(t)$ varies from a specified minimum to a specified maximum. For the results shown here, the minimum and maximum are separated by 0.2. For example, the first experiment looks at $b(t) \in [-0.9, -0.7]$ with $E[b(t)] = -0.8$ (mostly water), the second experiment looks at $b(t) \in [-0.7, -0.5]$ with $E[p(t)] = -0.6$, and so on. For each experiment three data sets were drawn from the true state, where each set has the same number of samples. The samples were drawn such that sensitivity and specificity of each data set was the same as in Scenario 1. Bootstrapping was also carried out in the same manner, so the statistics of the estimated sensitivity and specificity could be obtained.

The errors in the retrieved sensitivities and specificities are shown in Fig 2 as a function of the expected value of the estimated class imbalance, $E[\hat{b}]$. For the sensitivity the maximum error is for $\psi_3$, when the class imbalance is low ($E[\hat{b}]$=-0.793), while for the specificity the error is highest for

$\eta_3$, when the class imbalance is high ($E[\hat{b}]$=0.801). This is due to the fact that for those values of $E[\hat{b}]$ the denominators of equations (5) and (6) are close to zero for the sensitivity and specificity respectively, and because the term containing the class imbalance is multiplied by $\boldsymbol{v}$ and the element of $\boldsymbol{v}$ that is largest is that for the third data set (because it is ranked highest, i.e. it has the highest balanced accuracy). Therefore, the sensitivity of the denominator to small errors in $E[\hat{b}]$ is amplified more for this data set. For the results shown here, in all cases the errors in the retrieved $E[\hat{b}]$ were less than 0.01. For larger absolute values of $E[\hat{b}]$ the errors were higher, with an error in $E[\hat{b}]$ on the order of 20% approaching $E[\hat{b}] \pm 1$. These results were used to provide guidance on the interpretation of the estimates obtained using real data. For example, we expect higher errors in the proportion correct ice (sensitivity) when there is less ice in the dataset, and higher errors in the proportion correct water (specificity), when there is less water in the dataset.

## IV. Experiments with real data

### A. Study region and data

*1) Description of study region:* The study region considered here is the Gulf of Saint Lawrence (GSL). This is a seasonal ice zone where the Saint Lawrence river widens to an estuary and meets the Atlantic Ocean, situated along the east coast of Canada. The GSL is a busy shipping zone during the winter, and is the location of most ice-related ship besetments in Canada [5]. For this study we consider the period of January 17th 2014 to February 10th 2014. A daily ice chart of the study region for January 17th 2014 is shown in Fig 3 (top). This ice chart represents a manually generated representation of the stage of development of the ice cover prepared by an ice analyst using a variety of satellite and in-situ data. It can be seen that on this date the ice cover in the region was dominated with new, grey and grey-white ice, which are estimated to have ice thickness values less than 30 cm [1]. A time series of the daily average air temperature from a weather station located on Îles de la Madeline, near the middle of the study region, is shown in Fig 3 (bottom). Air temperatures varied from around freezing at the beginning of the study period, to approximately -15$^o$C toward the end. Further inspection of daily ice charts found that over this time the region transitioned from a thin and sparse ice cover, to a consolidated cover of thicker first-year ice. The study region and period were chosen because they represent freeze-up conditions in a first-year ice zone, which can be problematic for sea ice data sets derived from either passive or active microwave imagery. For passive microwave imagery, the concentration of thin ice is underestimated by sea ice concentration retrievals because the brightness temperature under these conditions cannot be distinguished from that of thicker ice at a lower ice concentration [12], in particular for retrieval algorithms using low frequency channels. For SAR imagery, when thin ice is smooth the backscatter is low and it is easily mistaken for calm open water, whereas thin ice covered with frost flowers has much higher backscatter and an be mistaken for thicker ice [28].

*2) Data:*

*Ice concentration from the ARTIST Sea Ice (ASI) algorithm*: Ice concentration calculated using the ASI algorithm [30] was used as the passive microwave sea ice concentration data set. The ice concentration in the ASI algorithm is calculated using data from the 89 GHz channels of the AMSR2 sensor and is available once per day on a polar stereographic grid with grid spacing of 3.125 km [3]. It can be downloaded from http://icdc.zmaw.de/seaiceconcentration_asi_amsre.html. It should be noted that spatial resolution of the data is limited by the instrument field of view for 89 GHz on the AMSR2 sensor, which is approximately 3 km $\times$ 5 km, and is also impacted by the use of lower frequency channels in the weather filters. ASI ice concentration has been compared with ice concentration from ship-based observations [30] and from Landsat images [33]. The errors were highest for areas of low ice concentration, and in areas covered by new ice, which is ice less than 10 cm in thickness. The underestimate for new ice, and more generally thin ice (thickness less than 30 cm) is typical of ice concentration estimates from passive microwave sensors [12]. The underestimation is lower for algorithms using the high frequency channels, such as the ASI algorithm. Since the study region contains large quantities of thin ice, the ASI ice concentration was considered an appropriate choice among the available passive microwave sea ice concentration retrieval products.

*Ice concentration calculated from SAR*: The second sea ice concentration data set is provided by an automated method that uses a convolutional neural network (CNN) to learn sea ice concentration from a set of SAR images. The CNN is described fully in [32], and is reviewed only briefly here. The method uses operational SAR imagery, which is acquired at 5.3 GHz (C band) in ScanSAR wide mode, with a swath range of 500 km and a nominal pixel spacing of 50 m. Spatial averaging is first applied to the images to reduce speckle noise. The CNN consists of a stack of alternating layers, with the layers alternating between convolutional filters and pooling. The top layer carries out linear regression, such that the output from the CNN is a continuous variable, the sea ice concentration. The filter weights in the CNN are learned iteratively from the difference between the ice concentration predicted by the method and that from image analysis charts. The CNN is trained using a set of images from the Beaufort Sea, north of Alaska, as well as a set acquired over the Gulf of St. Lawrence. The spatial resolution of the CNN output is variable, as the data are represented using convolutional filters with a variety of receptive fields. The smallest scale the CNN can represent is approximately 800m (twice the grid spacing used in generating the ice concentration), and the largest scale is approximately 4km (according to the scale of the largest convolutional filter used).

*Ice concentration from the coupled ice-ocean-atmosphere forecasting system*: The thired sea ice concentration data set is provided by ice concentration analyses from an operational coupled ice-ocean-atmosphere forecasting system over the Gulf of St. Lawrence, Canada [29]. The forecasting system

uses an ice-ocean model (Nucleus for European Modelling of the Ocean coupled with the Community Ice Code, or NEMO-CICE) that is coupled to an atmospheric model that is a regional version of Environment Canada's Global Environmental Model (GEM). The grid resolution is approximately $0.05^o \times 0.03^o$ in longitude and latitude. The system uses direct insertion of sea ice concentration from image analysis charts to assimilate observational sea ice data into the model state. Due to the fact that image analysis charts are based on SAR imagery, it is possible that the ice concentration analyses from the coupled forecasting system are correlated with the ice concentration from the CNN. To mitigate this the forecasts used in CTC are those available directly before the acquisition time of the SAR image that is used to calculate ice concentration. However, the coupled forecasting system would still have seen the CNN ice concentration from the previous SAR image acquisition, which means the errors in the forecast may not be independent from those from the CNN ice concentration. For example, consider that the image analysis does not contain an opening in the ice cover, and this persists through the forecasts, which are then used with an ice concentration from the CNN in CTC. If this ice concentration from the CNN also misses the same opening, this would be a correlated error. A similar situation could arise if the assimilated image analysis overestimates the extent of the marginal ice zone, and this situation persists through the forecasts. For the case considered here, such errors are not likely an issue because images acquired on sequential dates typically do not overlap the same geographic region. There are only two sets of sequential dates of those used in the analysis with overlap of imagery. In both cases, the overlap is small (less than 20% of the study region).

### B. Experimental Setup

The three sea ice concentration data sets contain 2D spatially distributed fields. To apply CTC to these data, for each fixed date, a single set of scores will be obtained for three sets of ice/water data over a spatial region. In this case, on each fixed date, the expected class imbalance over the spatial region will differ as the relative frequency of ice vs water changes as the ice season progresses. This is similar to the methodology in [21], where at each fixed location, a single set of rankings were obtained for three sets of freeze-thaw time series data. In that case, at each fixed location, the expected relative frequency of freeze vs thaw was expected to change with time due to seasonality.

To obtain a binary ice/water state for our data it is required to threshold the ice concentration, such that values above a specified value, $IC_{thresh}$ are considered ice, while those below $IC_{thresh}$, are considered water. Two values of $IC_{thresh}$ were considered $IC_{thresh} = 0.1$, and $IC_{thresh} = 0.3$, as they represent commonly used thresholds in ice operations and climate studies [22, 1]. For each date in the study period on which all three data sets are available (there is not a SAR image acquisition each day), the data sets are first collocated to the same spatial grid, which was chosen to have a spatial resolution of 4 km to be approximately consistent with the nominal spatial resolution of the ice-ocean model and the ASI sea ice concentration. Differences in scale between the data sets used in triple collocation can lead to representativity errors that can artificially increase or decrease the ranking of a dataset [34]. Here the scales of the datasets are relatively similar, and representativity errors are not likely a significant issue.

Using these collocated data sets, an estimate of $E[\hat{b}]$, $\psi$ and $\eta$ is made each day. For these experiments, because the samples used for each estimate of $E[\hat{b}]$ are distributed spatially, the inhomogeneity is in space, not time. However, the formulation is the same as that given in Section II, with $t$ replaced by a spatial variable. To screen out dates with unreliable results, we only used dates for which the number of samples available was greater than 1000, and the confidence interval for $E[\hat{b}]$ was less than 0.5. This results in values being estimated for 10 dates between Jan 17, 2014 and Feb 10, 2014, although collocated datasets were available on 18 days in the study period. To estimate a confidence interval for the reported statistics, bootstrapping was carried with replacement using 1000 replicates.

### C. Estimation b, $\psi$ and $\eta$ for sea ice data sets

The estimated class imbalance from the GSL dataset is shown in Fig 4. Moving through the study period (from left to right), $E[\hat{b}]$ indicates an ice cover progressing from one with significant open water (negative $E[\hat{b}]$ values), to one with more ice cover (positive $E[\hat{b}]$ values). This is consistent with expectations, since the GSL is freezing up at this time of year, with air temperatures decreasing and ice cover increasing from mid-January into February. The increase in $E[\hat{b}]$ is not monotonic as we progress through the ice season, largely due to the fact that each SAR image acquisition covers a different region, and therefore the region covered by the colocated data sets varies for each date in the study period. On some dates, such at January 31st and February 10th, the SAR image covers a large part of the central GSL, while on other dates, such as January 30, the acquisition covers the upper part of the estuary, which contains a larger open water fraction even when air temperatures are very cold due to the volume of fluid moving through that narrow region.

The estimated sensitivity and specificity on each date is shown in Fig 5 panels a) c) for the case when $IC_{thresh} = 0.1$ and Fig 5 panels b) d) for the case when $IC_{thresh} = 0.3$. It can be seen that for all dates up to January 31st, the sensitivity is the lowest for ASI, meaning that it is not able to accurately detect ice at this time. This is expected because at this time the region is covered with thin ice (less than 30 cm in thickness). It is known that passive microwave retrieval algorithms are biased toward predicting water for thin ice conditions [13]) when they use retrieval algorithms configured for Arctic conditions, which is the case here. In addition, it was found the weather filters were eliminating a significant portion of the ice cover during this period. As the ice season progresses the ice cover transitions from regions mainly consisting of thin, new ice, to thicker, first year ice, and the sensitivity of the ASI data set increases. On the other hand, the ASI data set has high values for specificity over almost all of the

dates when $IC_{thresh} = 0.3$. This is expected from equation (6) because the the mean value of the ASI classification will be low (negative) during this period due to the dominance of water, and if $E[\hat{b}]$ is also low, $\eta_i$ will be high. For both the ice-ocean model and the CNN, the sensitivity is generally high over the ice season. The specificities are lower, and in particular for the ice-ocean model the specificity is low when $IC_{thresh} = 0.1$. The impact of changing the ice concentration threshold is also illustrated in Table I. The scores for the CNN data set are not impacted as much by changes in the ice concentration threshold because these data have fewer points corresponding to intermediate ice concentrations than the data from ASI or the ice-ocean model.

For the ice-ocean model, change in specificity with $IC_{thresh}$ can be understood when we consider that the ice-ocean model has a more diffuse marginal ice zone as compared to ASI or the CNN output (see Fig 6). When $IC_{thresh} = 0.1$ is used the ice-covered points in the ice-ocean model data set with $IC > IC_{thresh}$ (for example a location in the ice-ocean model output that has $IC = 0.2$) would likely be occupied by water for the other two data sets (Fig 6 middle column). This means the covariance between the ice ocean model and each of the other two data sets would be low, leading to a low value of $v_i$ for the ice ocean model (see equations (5) and (6) for the impact on sensitivity and specificity). In spite of this dependence on $IC_{thresh}$, it can be seen in Fig 5 that on all dates, changing the ice concentration threshold did not change the ranking of the datasets (within the confidence intervals).

### *D. Evaluation of binary data sets using IMS*

To complement the evaluation using the triple collocation method, in this section we evaluate the accuracy of the datasets by directly comparing the three ice/water data sets to those from the interactive multisensor snow and ice mapping system (IMS). This is closer to the evaluation method used in sea ice operations [6, 17]. The IMS product is generated manually, taking into account a wide variety of data sources. For the version used here, available during the first few months of 2014, the main data sources used by the analysts would be from passive microwave data and optical imagery. Due to the extensive cloud cover in our particular study region for the period of interest, we expect the IMS data are heavily dependent on contributions from passive microwave sensors. However, we do not necessarily expect these data to be in good agreement with the ice/water states from ASI because IMS uses ice concentration based on the lower frequency channels of the Special Sensor Microwave Imager (SSMI) and Special Sensor Microwave Imager/Sounder (SSMI/S) sensors [11], which are much coarser in spatial resolution than the 89 GHz channel from AMSR2 used in the ASI algorithm (25 km gridded product as compared to 3.125 km).

| Data set | $\psi$ | $\eta$ | $\pi$ |
|---|---|---|---|
| CTC: $IC_{thresh} = 0.1$ | | | |
| ice-ocean model | 0.7941 | 0.5510 | 0.6728 |
| ASI | 0.7120 | 0.7592 | 0.7356 |
| CNN | 0.8185 | 0.7435 | 0.7810 |
| CTC: $IC_{thresh} = 0.3$ | | | |
| ice-ocean model | 0.7809 | 0.6169 | 0.6989 |
| ASI | 0.5798 | 0.7906 | 0.6852 |
| CNN | 0.8157 | 0.7015 | 0.7586 |
| IMS: $IC_{thresh} = 0.3$ | | | |
| ice-ocean model | 0.6952 | 0.8155 | 0.7182 |
| ASI | 0.4470 | 0.9091 | 0.6707 |
| CNN | 0.6165 | 0.8177 | 0.6990 |

Table I: Average balanced accuracy, sensitivity and specificity for each data set over the study period for two different ice concentration thresholds. Results are shown for both CTC (categorical triple collocation) and IMS (interactive multisensor snow and ice mapping system). Scores calculated using IMS utilize the same data points as for CTC.

In the comparison using IMS data, the proportion correct ice ($\psi_i$), proportion correct water ($\eta_i$) and balanced accuracy ($\pi_i$) were calculated by checking the agreement of each point classified as ice or water in the IMS analysis with that from the $i'th$ data set. These scores were obtained by applying $IC_{thresh} = 0.3$ to the ice concentration from each data set to convert it to a set of binary values. The value of 0.3 was chosen to be consistent with the analysis in the previous section. This threshold is slightly different with the threshold qualitatively applied to generate the ice/water points in IMS, which is 0.4. We note that when the analysis was done with $IC_{thresh} = 0.4$ the results were similar to $IC_{thresh} = 0.3$, with a slightly lower sensitivity of the ASI dataset.

It can be seen in Table I that the overall balanced accuracies are different from those from the CTC method. When IMS is used as the verification data, it is the ice-ocean model that has the highest balanced accuracy, followed by the CNN and ASI, whereas for the CTC method it is the CNN that has the highest balanced accuracy, followed by the ice-ocean model and ASI. Similar to the CTC ranking, the ASI data set has the lowest sensitivity and highest specificity of the three. The main difference is that when the IMS data are used the sensitivity of the model is much higher than the CNN, while the specificities of the two are much closer.

To understand the CTC and IMS comparison scores better, we show output from a given day (January 25th, 2014) in Fig 7. The data are shown here with the SAR HH image from the same date overlaid so the features of the ice cover can be visualized. In Fig 7 it can be seen that the ice concentration from the model and the CNN look more similar to each other than that from the ASI, and both would have a larger mean value than ASI (more ice), which explains why the sensitivity scores from CTC are higher for these two data sets on this date. We also compared the sensitivity and specificity scores for each data set from CTC with the proportion correct of ice and water from IMS. These scores are reported in Table II. The main difference between the two methods is in the proportion correct water (or specificity). These scores are high for ASI for both methods, and it can be seen that the water points in the IMS data are in good agreement with those in the ASI data, whereas the there is section of water in IMS that would be labelled as ice in the ice-ocean model and the CNN results. From comparison with the underlying SAR image it

appears this section may be incorrectly labelled by IMS, which could be due to the coarse spatial resolution of the input data, or due to the time difference between the IMS analysis and the SAR image acquisition, given the dynamic nature of the cover. We expect results would be different if a similar analysis was done with the more recent IMS product that incorporates SAR data. It can also be seen that the scales of information available in the IMS are somewhat similar to those from the ice-ocean model, which may be partially responsible for the higher overall balanced accuracy of this data set when IMS is used as the verification data.

| Data set | $\psi$ | $\eta$ | $\pi$ |
|---|---|---|---|
| January 25, 2014 | | | |
| CTC: $IC_{thresh} = 0.3$ | | | |
| ice-ocean-model | 0.941 | 0.908 | 0.925 |
| ASI | 0.346 | 0.928 | 0.662 |
| CNN | 0.945 | 0.742 | 0.844 |
| IMS: $IC_{thresh} = 0.3$ | | | |
| ice-ocean model | 0.810 | 0.559 | 0.685 |
| ASI | 0.323 | 0.935 | 0.629 |
| CNN | 0.838 | 0.404 | 0.621 |

Table II: Balanced accuracy, sensitivity and specificity for each data set for January 25, 2014. Results are shown for both CTC (categorical triple collocation) and IMS (interactive multisensor snow and ice mapping system). Scores calculated using IMS utilize the same data points as for CTC.

## V. Discussion

Results from using CTC in a seasonal ice zone covering a period from freeze-up to a thicker first-year ice cover show that either the ice-water states from the ice-ocean model or from the SAR data (via a CNN) are ranked first. We note the low ranking of the ASI ice concentration is not unexpected, and different results would be found in a different region with thicker or more consolidated ice, or different atmospheric conditions, which would mean a different response of the weather filters. Here the ASI algorithm is used as the ice concentration from passive microwave data. This algorithm uses weather filters that set the ice concentration to zero when the retrieved value is less than 15% [30]. Assuming a marginal ice zone exists on a given date, as we go from an ice concentration threshold of 10% to 30% when we binarize the ice concentration, we cross the 15% used by the weather filters. When we use a threshold of 30% in the binarization, points with ice concentration less than 30% are considered water. These water points will include the points in the ASI data that have been set to water by the weather filters, but may be in the marginal ice zone. This decreases the ranking of the ASI data, because if the points are considered ice in the other two datasets the covariance between these two other datasets may be large. In addition, there may be many more points in the ASI dataset with a water label when a threshold of 30% is used, as compared to the other two data sets, because of the weather filters. This decreases the mean of this data set (making it more negative). Together, these strongly decrease the sensitivity of the ASI data when going from a threshold of 10% to 30%. It would be interesting carry out a more in depth investigation of this issue, taking a critical look at the use of weather filters.

To visualize the ranking of the datasets, we have generated RGB plots (Fig 8). In these plots red indicates a pixel where for the given number of days there was a ranking at that pixel, the model was ranked first, while green and blue indicate where the ASI and CNN data were ranked first respectively. The lines in the image indicate the boundaries of the SAR image swaths. It can be seen in Fig 8a that when $IC_{thresh} = 0.1$ the CNN and ice-ocean model are generally first, with the CNN more dominant over the central part of the GSL and the ice ocean model performing better in the upper part of the estuary. When $IC_{thresh} = 0.3$ the rankings are more mixed, with the ASI data ranked first in some regions, which coincide with image acquisitions covering open water regions. For example, the band of green around $63^o$W in Fig 8b is due the image from January 17th, when the ASI data had a high ranking for specificity.

While the present study has examined a small region, in the future it would be interesting to apply this methodology to the broader Arctic, and extend the methodology to enable the use of more than three datasets. For the broader Arctic analysis could be carried out either looking at a time series at fixed location or by dividing the globe into regions to obtain statistics that could be used to calculate the sensitivity, specificity and balanced accuracy for each data set over each region. A comparison between various datasets would be useful toward selecting the best dataset, or merging datasets (see [10]) for an application of linear triple collocation for merging soil moisture datasets under the European Space Agency Climate Change Initiative for soil moisture). However, given that the estimates of sensitivity and specificity for CTC are not well defined when the study region is predominantly ice covered (eg., $b \approx 1$), such a study will require some consideration. For example, instead of examining ice/water states, it may be more fruitful to look at anomalies from a mean state, which would require a reformulation of the problem.

## VI. Conclusions

Results from this study show that CTC can be used to quantitatively rank binary data sets, and that the class imbalance can be determined and used to estimate scores for sensitivity and specificity. The accuracy of the estimated class imbalance, sensitivity and specificity are higher if the data are relatively balanced and there are sufficient samples. The required balance and number of samples are dependent on each other, and on the data itself. Qualitatively, in agreement with earlier studies, on dates when two data sets have similar scores, this indicates a similarity between the two, while on dates when all three have different score, there should be noticeable differences between all three data sets. This follows from the role of the covariance in estimating $\boldsymbol{v}$ and $\boldsymbol{\pi}$. The estimates of sensitivity and specificity provide more information than a ranking, and can more clearly show the agreement or disagreement between

various classifiers in terms of their estimates of each class. For the study here carried out during freeze-up, the results clearly indicate the challenge passive microwave data are experiencing in this region capturing thin ice, and its success in capturing water. We note that the results here are similar to, but not identical with, those obtained when a verification is done using IMS as the verification data set. The CTC method should be considered complementary information to such a verification, but we note that it is different. CTC provides information on agreement between classifiers, whereas when a data set is chosen as verification data, the scores are relative to that particular data set. Future work will consider the broader Arctic region, extension of the method to more than three datasets, and will compare results with the more recent IMS analyses, which use SAR in addition to data from newer sensors.

## VII. Acknowledgements

The authors would like to thank Environment and Climate Change Canada for providing the ice-ocean model output and the SAR images, the National Snow and Ice Data Center for providing the IMS data, and the University of Bremen for providing the ASI data.

## Appendix

To demonstrate that the expression for the covaraince tensor, introduced by [16], can be used for the case where the true state is non-stationary, we start by obtaining the same result as in [21] but using a slightly different method. To begin we write the expression for the covariance between two classifier outputs as

$$cov(X_i, X_j) = E[X_i X_j] - E[X_i]E[X_j] \tag{A1}$$

where $E[\cdot]$ denotes the expectation operator. As pointed out by [21] this expression does not take into account the fact that the data sets represent samples from a non-stationary state. For example, if the non-stationarity is with respect to time, the expectation of the classifier output is conditioned on time as, $E[X_i|t]$. Using the law of total expectation equation (A1) can be written as [4]

$$cov(X_i, X_j) = E[E[X_i X_j|t]] - E[E[X_i|t]]E[E[X_j|t]] \tag{A2}$$

We substitute into (A2) the expressions from [16] for $E[E[X_i X_j|t]], E[E[X_i|t]], E[E[X_j|t]]$ taking into account the time dependence for the non-stationary case. Hence, where in [16] they have for the stationary case,

$$E[X_i X_j] = p\psi_i\psi_j + (1-p)(1-\eta_i)(1-\eta_j), \tag{A3}$$

we follow [21] and condition on time, $t$, (i.e. $p(t) = Pr(T = 1|t)$),

$$E[X_i X_j|t] = p(t)\psi_i\psi_j + (1-p(t))(1-\eta_i)(1-\eta_j). \tag{A4}$$

Similar, for the mean of given classifier, we start with the result from [16]

$$E[X_i] = p\psi_i + (1-p)(1-\eta_i) \tag{A5}$$

and condition on $t$ to have

$$E[X_i|t] = p(t)\psi_i + (1-p(t))(1-\eta_i). \tag{A6}$$

Substitution of (A4) and (A6) into (A2), applying the expectation operator, and simplifying, yields

$$\begin{aligned} cov(X_i, X_j) =& E[p(t)]\psi_i\psi_j + (1-E[p(t)])(1-\eta_i)(1-\eta_j) - \\ & (E[p(t)]\psi_i + (1-E[p(t)])(1-\eta_i)) \\ & (E[p(t)]\psi_j + (1-E[p(t)])(1-\eta_j)) \end{aligned} \tag{A7}$$

Expanding and simplifying, using $\pi_i = \frac{\psi_i+\eta_i}{2}$ we arrive at

$$cov(X_i, Y_j) = E[p(t)](1-E[p(t)])\,(2\pi_i - 1)(2\pi_j - 1) \tag{A8}$$

which is consistent with [21] with the exception of a factor of $4$. This is due to the scaling introduced in [16], where the classifier outputs were expressed as $\tilde{X}_i = (X_i + 1)/2$ which means the covariance arrived at using the expressions in [16] is smaller than that in [21] by a factor of 4.

For the covariance tensor $cov(X_i, X_j, X_k)$, we start in the same way as for the covariance matrix

$$\begin{aligned} cov(X_i, X_j, X_k) = & E[X_i, X_j, X_j] - E[X_i]E[X_j, X_k] \\ & - E[X_j]E[X_i, X_k] - E[X_k]E[X_i, X_j] \\ & + 2E[X_i]E[X_j]E[X_k] \end{aligned} \tag{A9}$$

Again, using the law of total expectation [4],

$$\begin{aligned} cov(X_i, X_j, X_i) = & E[E[X_i, X_j, X_k|t]] \\ & - E[E[X_i|t]]E[E[X_j, X_k|t]] \\ & - E[E[X_j|t]]E[E[X_i, X_k|t]] \\ & - E[E[X_k|t]]E[E[X_i, X_j|t]] \\ & + 2E[E[X_i|t]]E[E[X_j|t]]E[E[X_k|t]]. \end{aligned} \tag{A10}$$

For the first term, we start with (following [16])

$$E[X_i, X_j, X_k] = p\psi_i\psi_j\psi_k + (1-p)(1-\eta_i)(1-\eta_j)(1-\eta_k) \tag{A11}$$

Conditioning on $t$

$$\begin{aligned} E[X_i, X_j, X_k|t] = & p(t)\psi_i\psi_j\psi_k \\ & + (1-p(t))(1-\eta_i)(1-\eta_j)(1-\eta_k) \end{aligned} \tag{A12}$$

Substitution of (A12),(A4) and (A6) into (A10), taking expectations, simplifying, rescaling and using $p(t) = \frac{1+b(t)}{2}$, yields,

$$T_{ijk} = -2E[b(t)](1-E[b(t)]^2)(2\pi_i - 1)(2\pi_j - 1)(2\pi_k - 1) \tag{A13}$$

where $T_{ijk}$ is the $i, j, k$ component of the covariance tensor and $i \neq j \neq k$. Substitution of $v_i = \sqrt{1-(E[b(t)])^2}(2\pi_i - 1)$ and similarly for $j$ and $k$, yields

$$T_{ijk} = -\left(2E[b(t)]/\sqrt{1-E[b(t)]^2}\right) v_i v_j v_k \tag{A14}$$

**K. Andrea Scott** 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.

## REFERENCES


[1] MANICE Manual of Standard Procedure for Observing and Reporting Ice Conditions. Issued by the Meteorological Service of Canada, 2005.

[2] Y. Batrak and M. Müller. Atmospheric response to kilometer-scale changes in sea ice concentration within the marginal ice zone. *Geophysical Research Letters*, 45:6702–6709, 2018.

[3] A. Beitsch, L. Kaleschke, and S. Kern. Investigating high-resolution AMSR2 sea ice concentrations during the February 2013 fracture event in the Beaufort Sea. *Remote Sensing*, 6(5):3841–3856, 2014.

[4] D.P. Bertsekas and J.N. Tsitsiklis. *Introduction to Probability*. Athena Scientific, 2008.

[5] T. Browne, D. Fowler, I. Kubat, D. Watson, and M. Sayed. An examination of the besetting of the MV Berge Atlantic. In *Proceedings of the twenty-eighth International Ocean and Polar Engineering Conference*, 2018.

[6] M. Buehner, A. Caya, L. Pogson, T. Carrieres, and P. Pestieau. A new Environment Canada regional ice analysis system. *Atmosphere-Ocean*, 51(1):18–34, 2013.

[7] S. Caires and A. Sterl. Validation of ocean wind and wave data using triple collocation. *Journal of Geophysical Research, Oceans*, 108(C3):3098, 2003.

[8] D.J. Cavalieri and C.L. Parkinson. Arctic sea ice variability and trends. *Cryosphere*, 6:881–889, 2012.

[9] L. Giustarini, R. Hostache, D. Kavetski, M. Chini, G. Corato, S. Schlaffer, and P. Matgen. Probabilistic flood mapping using synthetic aperture radar data. *IEEE Transactions on Geoscience and Remote Sensing*, 54:6958–6969, 2016.

[10] A. Gruber, W.A. Dorigo, W. Crow, and W. Wagner. Triple collocation-based merging of satellite soil moisture retrievals. *IEEE Transactions on Geoscience and Remote Sensing*, 55(12):6780–6792, 2017.

[11] S.R. Helfrich, D. McNamara, B.H. Ramsay, T. Baldwin, and T. Kasheta. Enhancements to, and forthcoming developments in the Interactive Multisensor Snow and Ice Mapping System. *Hydrological Processes*, 21:1576–1586, 2007.

[12] N. Ivanova, L.T. Pedersen, and R. Tonboe. D2.5 Product Validation and Algorithm Selection Report (PVASR). Sea Ice Concentration. Technical Report SICCI-PVSAR Version 1.1, European Space Agency, May 2013.

[13] N. Ivanova, L.T. Pedersen, R.T. Tonboe, S. Kern, T. Lavergne, A. Sorensen, R. Saldo, G. Dybkjaer, L. Brucker, and M. Shokr. Inter-comparison and evaluation of sea ice algorithms: towards further identification of challenges and optimal approach using passive microwave observations. *The Cryosphere*, 9:1797–1817, 2015.

[14] 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.

[15] A. Jaffe, E. Fetaya, B. Nadler, T. Jiang, and Y. Kluger. Unsupervised ensemble learning with dependent classifiers. *Proceedings of Machine Learning Research*, 51:351–360, 2016.

[16] A. Jaffe, B. Nadler, and Y. Kluger. Estimating the accuracies of multiple classifiers without labeled data. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, volume 38, pages 407–415, 2015.

[17] 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(10):5755–5769, 2017.

[18] 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.

[19] Y. Liu, J. Key, and R. Mahoney. Sea and freshwater ice concentration from VIIRS on Suomi NPP and the Future JPSS Satellite. *Remote Sensing*, 8, 2016.

[20] H. Lyu, K.A. McColl, X. Li, C. Derksen, A. Berg, A. Black, E. Euskirchen, M. Loranty, J. Pulliainen, K. Rautiainen, T. Rowlandson, A. Roy, A. Royer, A. Langlois, J. Stephens, H. Lu, and D. Entekhabi. Validation of SMAP freeze/thaw product using categorical triple collocation. *Remote Sensing of Environment*, 205:329–337, 2018.

[21] K.A. McColl, A. Roy, C. Derksen, A.G. Konings, S.H. Alemohammed, and D. Entekhabi. Triple collocation for binary and categorical variables: Application to validating landscape freeze/thaw retrievals. *Remote Sensing of Environment*, 176:31–42, 2016.

[22] W. Meier and J. Stroeve. Comparison of sea-ice extent and ice-edge location estimates from passive microwave and enhanced-resolution scatterometer data. *Annals of Glaciology*, 48:65–70, 2008.

[23] J. Murfitt, L.C. Brown, and S.E.L Howell. Evaluating RADARSAT-2 for monitoring of lake ice phenology events in mid-latitudes. *Remote Sensing*, 10, 2018.

[24] NSIDC: National Snow and Ice Data Center, Boulder Colorado USA. IMS Daily Northern Hemisphere Snow and Ice Analysis at 1km, 4km and 24km Resolutions, 2008. Accessed:March 10th 2019.

[25] F. Parisi, F. Strino, B. Nadler, and Y. Kluger. Ranking and combining multiple predictors without labeled data. *PNAS*, 111(4):1253–1258, 2014.

[26] D.K. Perovich. The optical properties of sea ice. Technical Report Monograph 96-1, US Army Cold Regions

Research Laboratory, 1996.

[27] K. Rittger, T.H. Painter, and J. Dozier. Assessment of methods for mapping snow cover from MODIS. *Advances in Water Resources*, 51:367–380, 2013.

[28] M. Shokr and N. Sinha. *Sea Ice: Physics, Mechanics, and Remote Sensing*. Geophysical monograph. American Geophysical Union, 2015.

[29] G.C. Smith, F. Roy, and B. Brasnett. Evaluation of an operational ice-ocean analysis and forecast system for the Gulf of St Lawrence. *Quarterly Journal of the Royal Meteorological Society*, 139:419–433, 2013.

[30] G Spreen, L. Kaleschke, and G. Heygster. Sea ice remote sensing using AMSR-E 89 GHz channels. *Journal of Geophysical Research*, 113(C2):C02303, 2008.

[31] A. Stoffelen. Toward the true near-surface wind speed: Error modelling and calibration using triple collocation. *Journal of Geophysical Research*, 103(C4):7755–7766, 1998.

[32] L. Wang, K.A. Scott, L. Xu, and D.A. Clausi. Sea ice concentration during melt from dual-pol SAR scenes using deep convolutional neural networks: A case study. *IEEE Transactions on Geoscience and Remote Sensing*, 58:1–10, 2016.

[33] H. Wiebe, G. Heygster, and T. Markus. Comparison of the ASI ice concentration algorithm with Landsat-7 ETM+ and SAR imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 47(5):3008–3015, 2009.

[34] T. Yilmaz and W. Crow. Evaluation of assumptions in soil moisture triple collocation analysis. *Journal of Hydrometeorology*, 15:1293–1302, 2014.

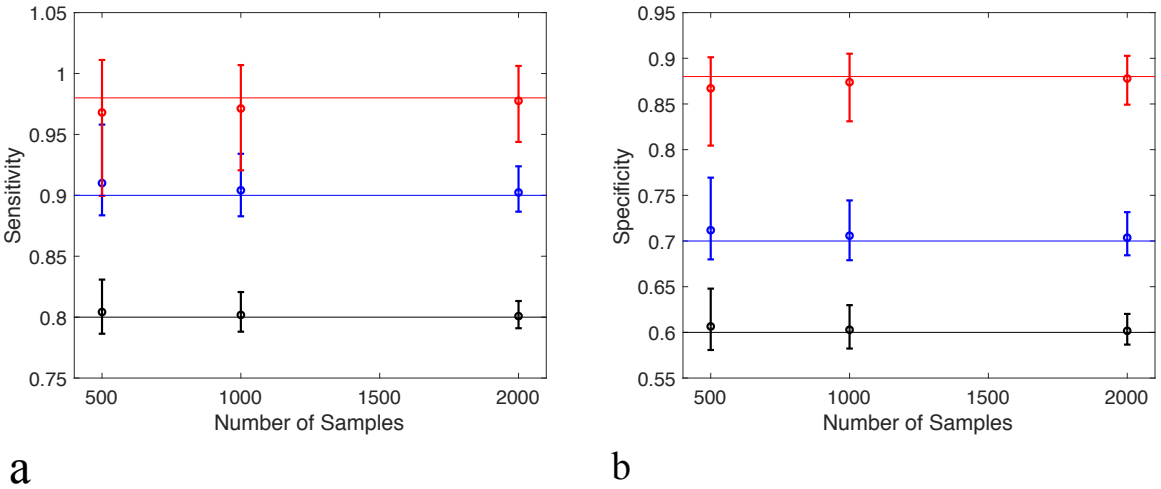


Figure 1: The estimated sensitivity (a) and specificity (b) for each of the three data sets as a function of the number of samples in the dataset for scenario 1, $E[b] = 0$. Set 1, black, Set 2, blue, Set 3, red. The dots indicate the mean values for estimated sensitivity and specificity, lines indicate the true values, and bars indicate the width of the 90% confidence interval.

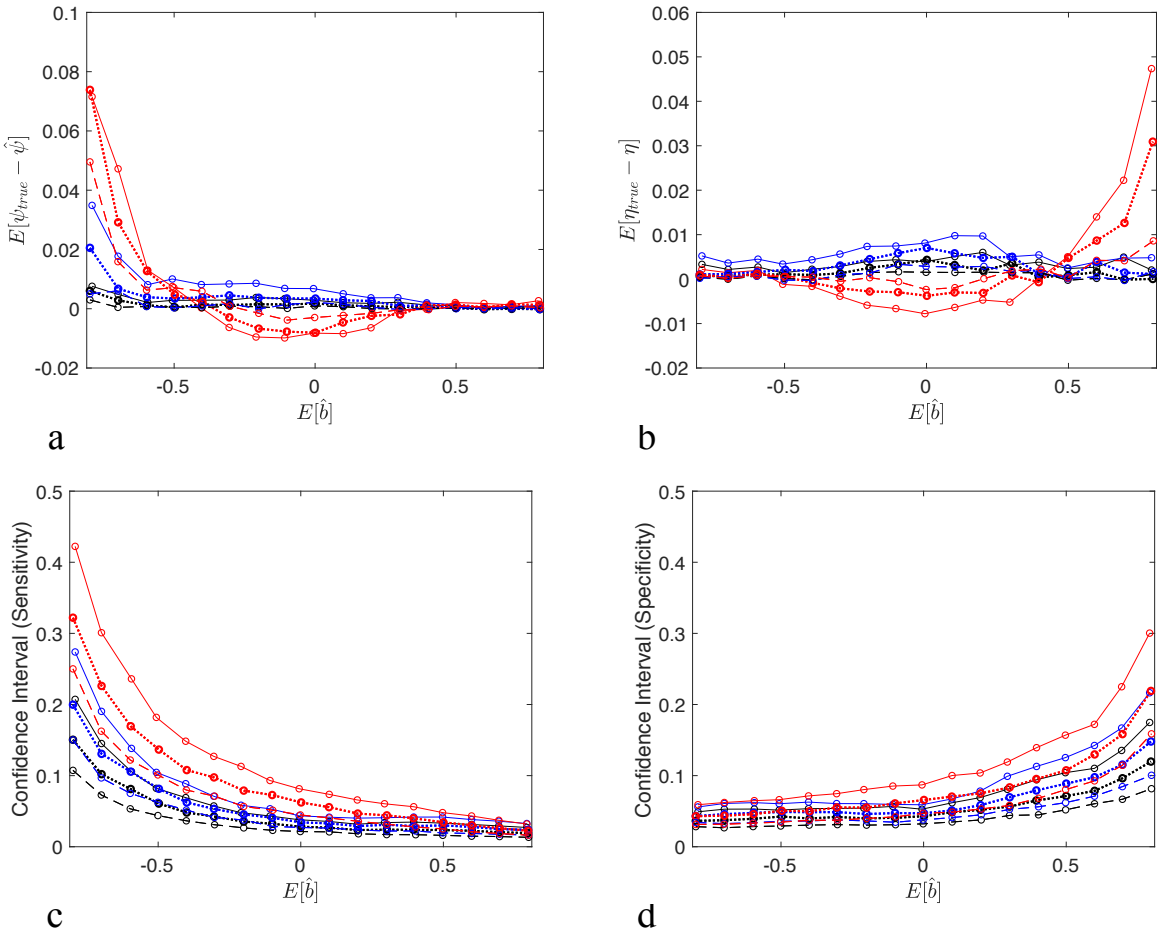


Figure 2: The error in sensitivity (a) and specificity (b) and 90% confidence interval for (c) sensitivity and (d) specificity for each of the three datasets as a function of the estimated balance for scenario 2, $E[b] \neq 0$. Set 1, black, Set 2, blue, Set 3, red. Solid lines are $n = 500$, dotted lines are $n = 1000$ and dashed lines are $n = 2000$.

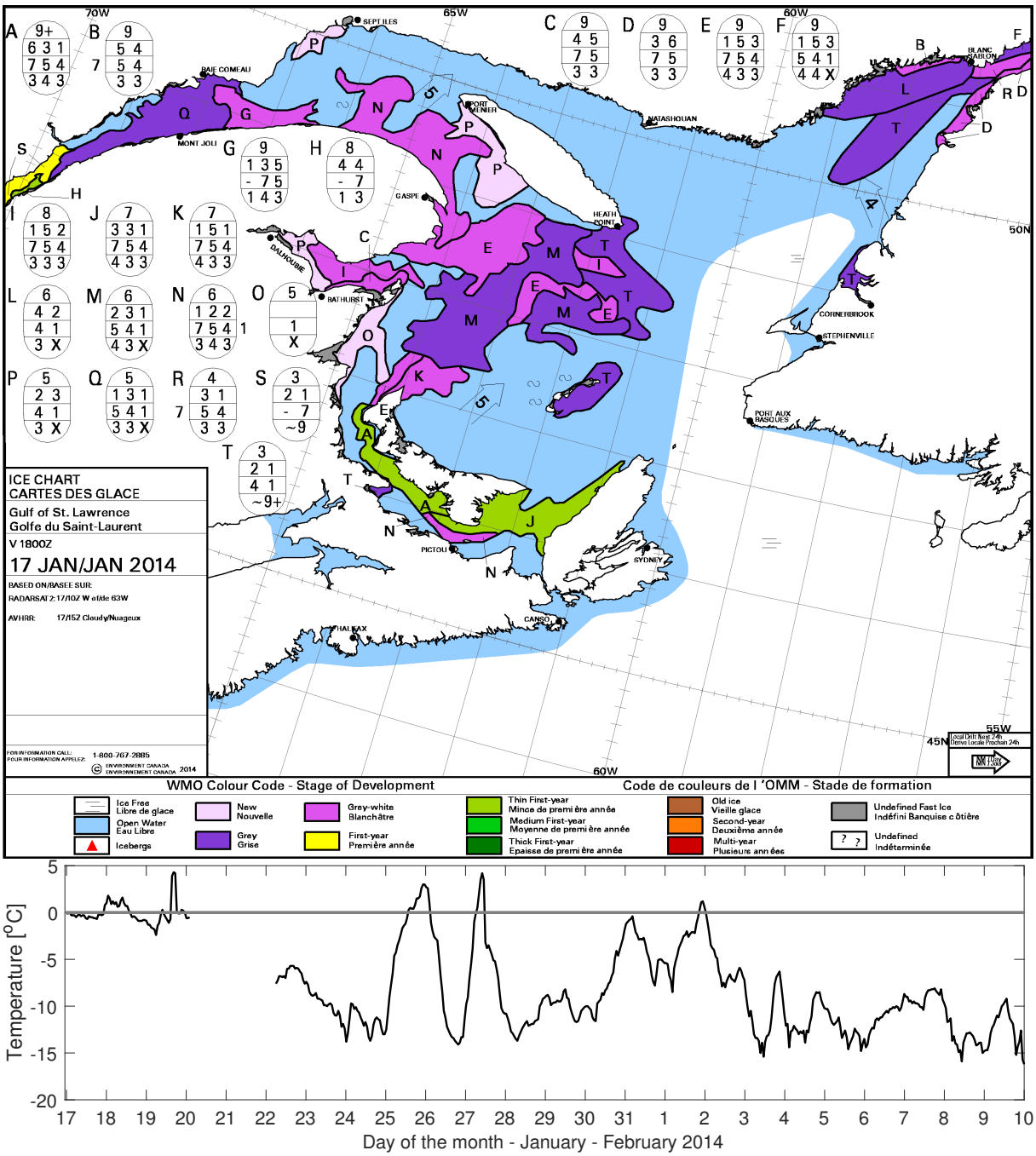


Figure 3: Top, Ice type information for the daily ice chart produced by the Canadian Ice Service on January 17th, 2014. It can be seen that on this date there is significant coverage of new ice, grey ice and grey-white ice, which corresponds to an estimated range of ice thicknesses less than 30cm [1] . Bottom, air temperature over the study period recorded at the weather station on Îles de la Madeline. The location of the weather station is shown by the red dot on the map.

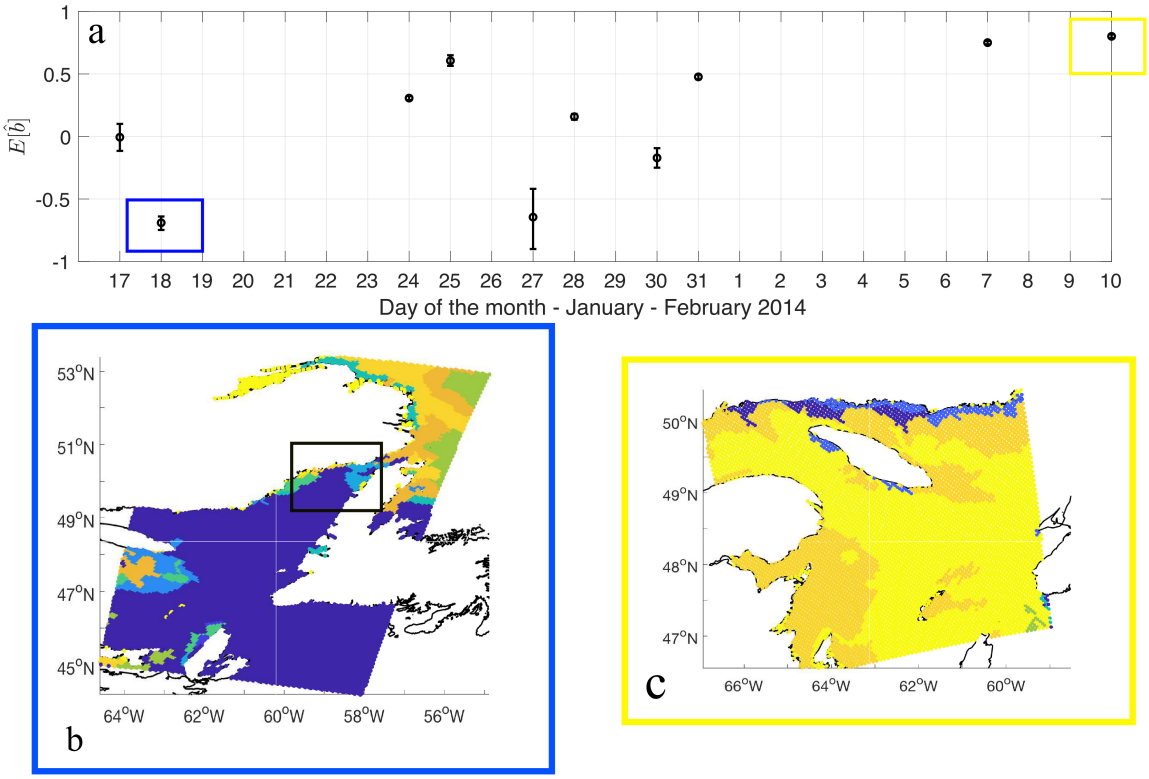


Figure 4: Class imbalance, panel a, for the case where $IC_{thresh} = 0.1$. Panels b and c are ice concentration maps from image analysis charts showing the ice concentration for the given day, blue is open water and yellow is ice concentration of 100%. Panel b is from January 18th 2014 (only region of image in GSL south of $51^o N$ is considered in the calculations done here) and panel c is from February 10th 2014. It can be seen that the case where the class imbalance is more negative corresponds to more open water in the image analysis chart, whereas the one where it is more positive corresponds to more ice cover, as expected.

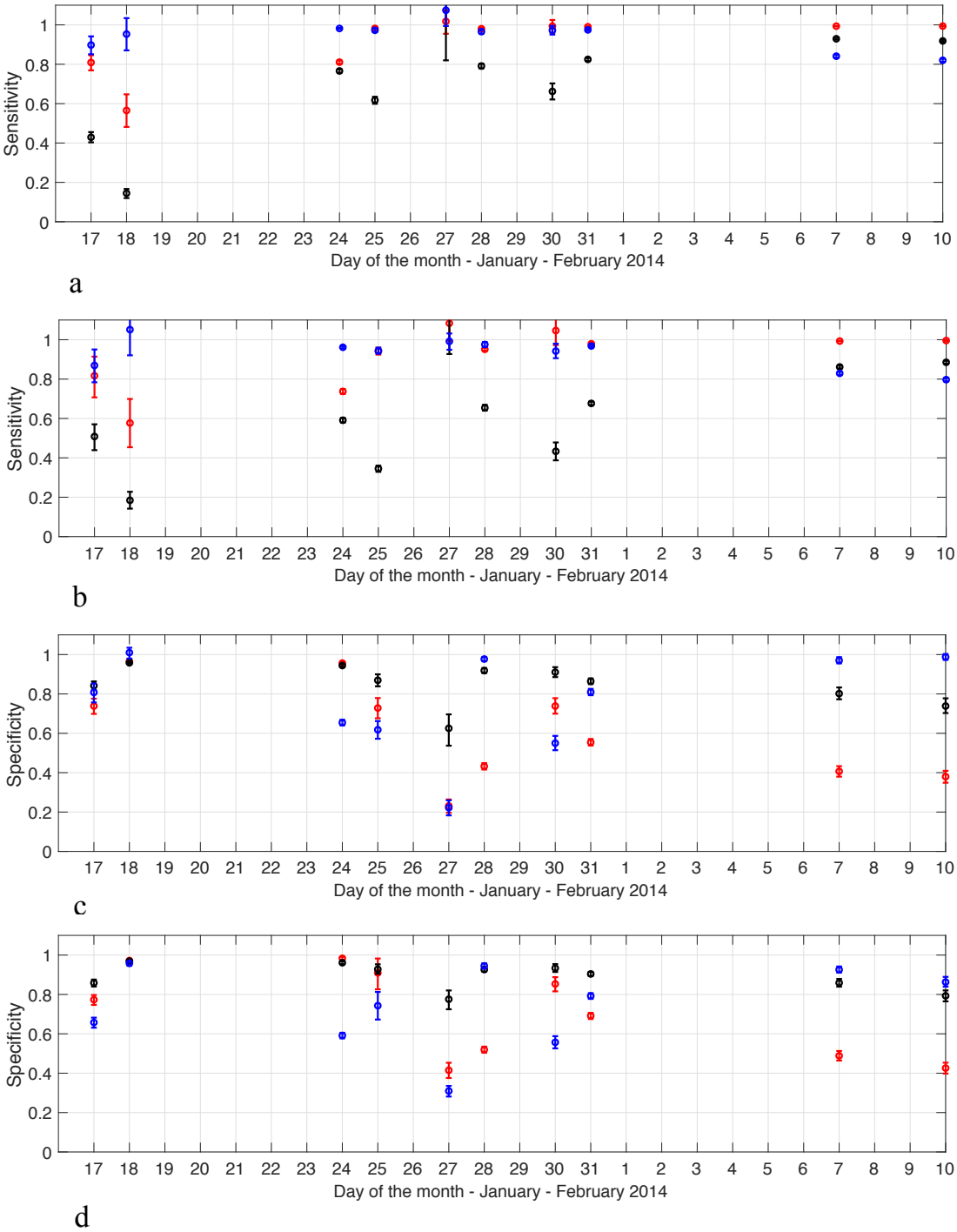


Figure 5: Estimated sensitivity and specificity of each data set for the Gulf of Saint Lawrence over the study period. Panels a) b) are the sensitivity and c) d) are the specificity. To show the impact of ice concentration threshold, a) c) $IC_{thresh} = 0.1$ and b) d) $IC_{thresh} = 0.3$. Red, ice-ocean model; black ASI, blue, CNN. Bars indicate the width of the 95% confidence interval.

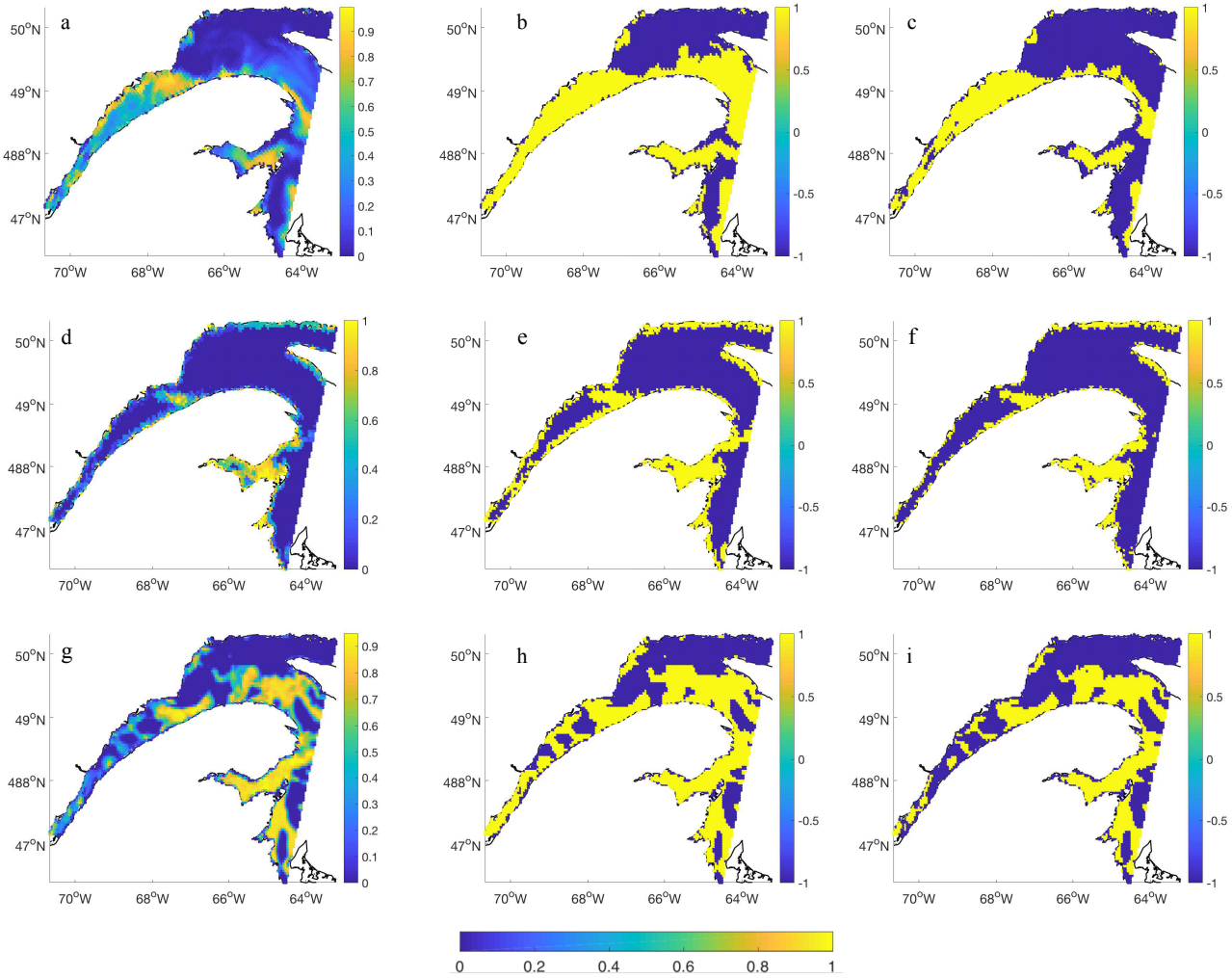


Figure 6: Sea ice concentration data sets. a) d) g), and thresholded sea ice concentration using b) e) h) $IC_{thresh} = 0.1$, c) f) i) $IC_{thresh} = 0.3$. Top row is for the ice-ocean model, middle row is for ASI and bottom row is for the CNN.

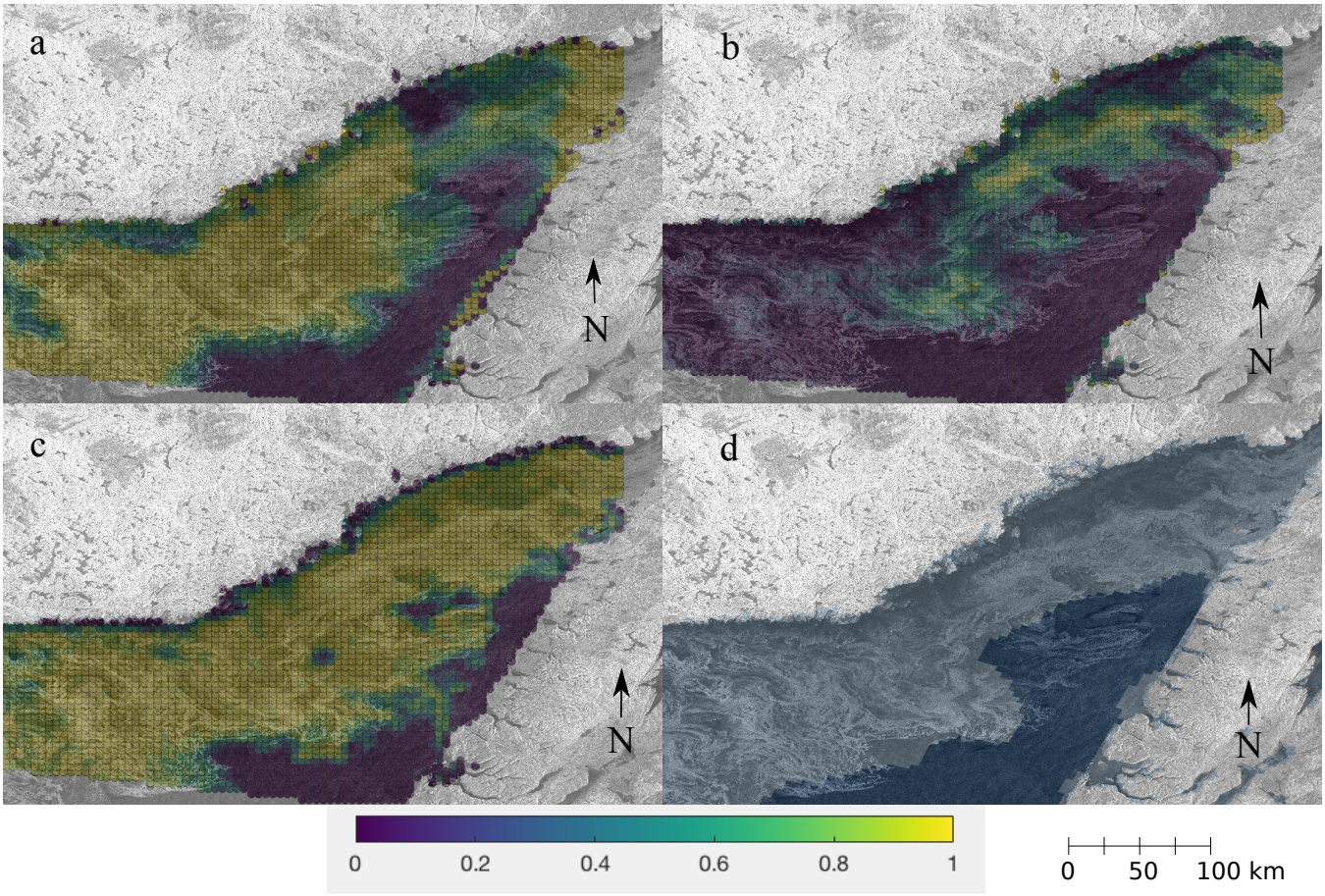


Figure 7: Ice concentration from the a) ice-ocean model b) ASI c) CNN data for January 25th, 2014. The region shown is corresponds to the black box in Fig 4b. Colorbar for ice concentration is shown at the bottom of the figure. The IMS ice/water labels for this date are also shown in panel d. For the IMS, dark blue is water and light blue is ice. Land is grey. All data are overlaid on the HH SAR image acquired for this date. Note the CNN is trained to estimate ice concentration from SAR imagery, so we expect good agreement between these two (panel c).

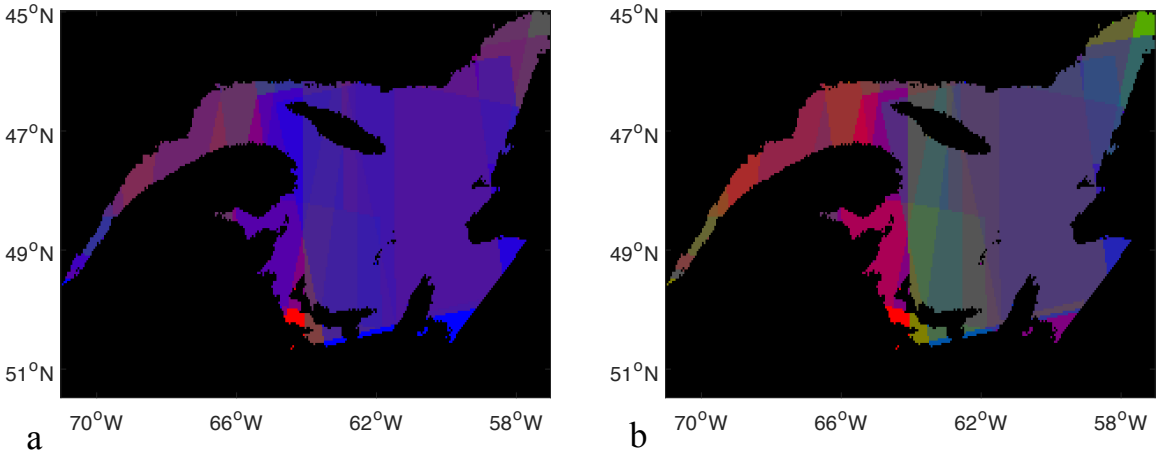


Figure 8: RGB plots indicating the fraction of the dates for which a given data set is ranked first. Red indicates the ice ocean model, blue indicates the CNN and green indicates ASI. Panel (a), $IC_{thresh} = 0.1$ and panel (b) $IC_{thresh} = 0.3$.