Data-driven Regularization and Uncertainty Estimation to Improve Sea Ice Data Assimilation

by

Nazanin Asadi

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2019

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:
Dr. Juha Karvonen
Senior Research Scientist
Finnish Meteorological Institute

Supervisors:
Dr. Katharine Andrea Scott
Assistant Professor, Systems Design Engineering
University of Waterloo

Dr. David A. Clausi
Professor, Systems Design Engineering
University of Waterloo

Internal Members:
Dr. Paul Fieguth
Professor, Systems Design Engineering
University of Waterloo

Dr. Nasser Lashgarian Azad
Associate Professor, Systems Design Engineering
University of Waterloo

Internal-External Member:
Dr. Ellsworth LeDrew
Professor, Geography and Environmental Management
University of Waterloo

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Content from 1 paper is used in this thesis. I was a co-author with major contributions on the design, development, evaluation and writing of the papers material.

N. Asadi, K. A. Scott, and D. A. Clausi. ”Data fusion and data assimilation of ice thickness observations using a regularisation framework.” *Tellus A: Dynamic Meteorology and Oceanography*, DOI:10.1080/16000870.2018.1564487: 2019.

This paper is incorporated in Chapter 3 of this thesis.

Abstract

Accurate estimates of sea ice conditions such as ice thickness and ice concentration in the ice-covered regions are critical for shipping activities, ice operations and weather forecasting. The need for this information has increased due to the recent record of decline in Arctic ice extent and thinning of the ice cover, which has resulted in more shipping activities and climate studies. Despite the extensive studies and progress to improve the quality of sea ice forecasts from prognostic models, there is still significant room for improvement. For example, ice-ocean models have difficulty estimating the ice thickness distribution accurately. To help improve model forecasts, data assimilation is used to combine observational data with model forecasts and produce more accurate estimates.

The assimilation of ice thickness observations, compared to other ice parameters such as ice concentration, is still relatively unexplored since the satellite-based ice thickness observations have only recently become common. Also, preserving sharp features of ice cover, such as leads and ridges, can be difficult, due to the spatial correlations in the background error covariance matrices. At the same time, the current ice concentration assimilation systems do not directly assimilate high resolution sea ice information from synthetic aperture radar (SAR), even though they are the main source of information for operational production of ice chart products at the Canadian Ice Service. The key challenge in SAR data assimilation is automating the interpretation of SAR images.

To address the problem of assimilating ice thickness observations while preserving sharp features, two different objective functions are studied. One with a conventional l_2-norm and one imposing an additional l_1-norm on the derivative of the ice thickness state estimate as a sparse regularization. The latter is motivated by analysis of high resolution ice thickness observations derived from an airborne electromagnetic sensor demonstrating the sparsity of the ice thickness in the derivative domain. The data fusion and data assimilation experiments are performed over a wide range of background and observation error correlation length scales. Results demonstrate the superiority of using a combined l_1-l_2 regularization framework especially when the background error correlation length scale was relatively short (approximately five times the analysis grid spacing).

The problem of automated information retrieval from SAR images has been explored in a problem of ice/water classification. The selected classification approach takes advantage of neural networks to produce results comparable to a previous study using logistic regression. The employed dataset in both studies is a comprehensive dataset consisting of 15405 SAR images over a seven year period, covering all months and different locations. In addition, recent neural network uncertainty estimation approaches are employed to estimate the uncertainty associated with the classification of ice/water labels, which was not explored in this problem domain previously. These predicted uncertainties can improve the automated classification process by identifying regions in the predictions that should be checked manually by an analyst.

Acknowledgements

First, I would like to express my sincere gratitude to my supervisors Professor Andrea Scott and Professor David Clausi for their guidance, advice, and moral support for the past five years. I have been extremely lucky to have supervisors who cared so much about my work, and who responded to my questions and queries so promptly.

I wish to thank my doctoral committee members, Prof. Paul Fieguth, Prof. Nasser Lashgarian Azad, Prof. Ellsworth LeDrew, and Prof. Juha Karvonen for their valuable comments and suggestions.

I would sincerely like to thank Professor Christian Haas for the use of the AEM data in Chapter 3 of this thesis. Also, many thanks to Dr. Alexander Komarov from the Environment and Climate Change Canada for kindly providing the ice/water database used in Chapters 4 and 5.

Special thanks to my dear friends in Toronto and Waterloo for their kind support and understanding during this period. I would specifically like to thank Navid, Parisa, Akram, Shervin, Peyman, Shima, Ala, Hengameh, Mohammad, and Sanaz for all the fun we had and the wonderful times we shared. I would also like to thank my friends and office-mates Amir and Mehdi for all the help and support specially through the past year.

Lastly, I would like to thank my dear family for their continuous and unparalleled love, help and support. I am forever indebted to my parents, Mahboubeh and Abbas, for pushing me to be the best that I can be and giving me the opportunities and experiences that have made me who I am. I am grateful to my brother, Payam, whom offered invaluable support and humor over the years and has been a light on my academic path. And most of all, I am grateful to the love of my life, Keyvan, for everything! This journey would have not been possible without you. *”You gave me a shoulder when I needed it; You showed me love when I wasn’t feeling it; You helped me fight when I was giving in; And you made me laugh when I was losing it”.*

Dedication

To my loved ones, my beloved parents, my dear husband, and my brother.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Problem Statement

Sea ice covers a large area of the ocean surface and plays an important role in ship navigation, fishing industries, oil exploration and construction operations, and global climate change. Sea ice extent and thickness are also important indicators of global warming. The remote locations and extreme climate of ice covered regions have made sea ice studies difficult. Initially, prior to the 1900s, sea ice observations were collected by people who lived in these regions. Later, ships and submarines were used to gather sea ice information. However, the provided information was limited to a focused area. Nowadays, satellite remote sensing provides the main source of observations for sea ice applications.

According to the recent records of sea ice concentration and thickness, there has been a significant decline in Arctic ice extent and thinning of the ice cover over the past two decades [98, 107]. The continuation of this trend facilitates marine transportation across the ice covered regions [44]. However, the safety of navigation in ice covered regions depends in part on the availability of accurate estimates of the small scale details of ice concentration, ice thickness and ice pressure. Ships will tend to navigate through narrow openings in the ice cover, and if there are none present they will attempt to find a route where the ice appears to be thin and/or undeformed [94]. Sudden changes in ice thickness (e.g., deformation ridges) and ice pressure play a significant role in vessel besetting events [81]. Given these information, it is required to have reliable sea ice forecasts over all regions and different time-scales.

Ideally, sea ice forecasts would be made in a similar manner to weather forecasts, using data assimilation [12]. Data assimilation is a method for combining available observations with a background state from a numerical forecast model, to find the best estimate of the current state of the system, which is called the analysis [83]. The analysis is then used to initialize the forecast model to run a short-term forecast, the output of which is used as the background state for the next data assimilation cycle. Using a data assimilation system would allow physically-based forecasts to be made of features in the ice cover that cannot be directly observed, such as regions of high ice pressure.

Data assimilation can also be considered an inverse problem where the objective is to find the best state, $\mathbf{x} \in \mathbb{R}^n$, that satisfies the relationship $\mathbf{y} = H(\mathbf{x}) + \epsilon$, where $\mathbf{y} \in \mathbb{R}^m$ represents the available observations, $H : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is the observation operator, which maps the state vector to the observation space and can be a physical or mathematical model, and $\epsilon \in \mathbb{R}^m$ is the observation error. The standard objective function in an inverse problem is defined as the l2-norm of the misfit between the observations and model state as $\|\mathbf{y} - H(\mathbf{x})\|_2^2$. Since, usually there are not enough observations available (i.e., the dimension of $\mathbf{y}$ is less than $\mathbf{x}$) and/or the observations have errors, the data assimilation problem may be very ill-posed. Regularization approaches are proposed to address this problem by adding regularization terms to the standard objective function.

Environment and Climate Change Canada (ECCC) has developed the Regional Ice-Ocean Prediction System (RIOPS) for operational short-range ice forecasting [89]. Even though high resolution synthetic aperture radar (SAR) observations currently provide the most reliable ice information, the assimilation process of RIOPS does not directly use these data. Instead, they actually use daily ice charts and image analyses products of Canadian Ice Service (CIS) manually provided based on SAR observation. In addition, they rely on ice information provided by optical sensor and low resolution passive microwave and scatterometer sensors. Thereby, the application of SAR images in the operational sea ice assimilation systems is limited to providing manual ice charts by CIS. The direct assimilation of automated extracted SAR information is still an open problem to address.

1.2 Challenges

Dealing with the observations is a major yet challenging task in the assimilation of sea ice observations. The major challenges of these tasks that are further addressed in this thesis are:

- Data assimilation schemes involve an interpolation of observational information to a model grid. To carry out this interpolation, certain criteria should generally be satisfied, such as smoothness of the underlying fields [95, 129]. This poses a specific problem for sea ice because even though at large scales sea ice seems to be a nonrigid continuum, at small scales it includes important sharp features or discontinuities [125]. These sharp features can be difficult to preserve in data fusion and data assimilation due to the spatial correlations in the background error covariance matrices. In addition, the current data assimilation systems assume constant and uncorrelated observation errors.

- The practical implementation of SAR-derived sea ice information, as the current primary source of information, in the sea ice data assimilation systems is still limited. Part of that limitation is due to the difficulty of automated information extraction from SAR images that forces the scientists to manually provide the products. This issue has been caused by the overlap of backscatter signatures from different ice types with open water. In addition, most of the current ice/water classification studies use a specific geographical area for training [62, 108, 2]. This is in contrast to most sea ice data assimilation systems, which cover larger regions.

- Another reason for limited utilization of SAR-derived sea ice information is the fact such information does not typically include any uncertainty estimates or confidence levels in such retrievals. While there are a few studies on uncertainty estimation of sea ice retrievals, for the task of ice/water classification there are no known published uncertainty estimates. Having the uncertainty of observations can provide additional insight for choosing useful observations to assimilate.

1.3 Objectives and Contributions

The objectives of this thesis target the problem of assimilating sea ice observations via the following three contributions.

1. *Demonstrating the existence of sparsity in sea ice thickness observations in the spatial derivative domain using both airborne survey data, and submarine sonar data, which is described in Chapter 3.* The sparsity information has been further used to show its impact on improving the quality of sea ice estimates when an additional regularization term is imposed on the cost function of the data fusion and data assimilation problem. The data assimilation experiments discussed in Chapter 3 not only examine the impact of background error correlations on the analyses, but also the impact of observation error correlations.

2. *Improving the quality of automated ice/water classification of RADARSAT-2 dual-polarization HH-HV images using a neural network approach, which is described in Chapter 4.* The experiments are designed to evaluate the classification performance considering different network sizes and feature inputs.

3. *Estimating the uncertainty of ice/water classifications of Chapter 4 by modifying the proposed neural network based on recent advances in neural networks, which is described and assessed in Chapter 5.* The neural network based uncertainty approaches include uncertainty induced by model parameters in addition to uncertainty induced by dataset features. These uncertainties have the potential to provide observation errors for data assimilation purposes.

1.4 Thesis Outline

The remainder of the thesis is organized as follows.

In Chapter 2, the essential background material required for understanding and motivating this research are provided. These include the overview of data assimilation and regularization, sea ice observations as inputs of sea ice data assimilation systems and some approaches on automated ice and water detection from SAR.

In chapter 3, the use of l_1-norm regularization when the data assimilation state exhibits sparsity in the actual or transformed domain is investigated in a sea ice assimilation problem. The experiments are conducted using real ice thickness observations and a 1-D sea ice model.

In chapter 4, a neural network approach is utilized to classify a comprehensive database of SAR image features into ice and open water.

The classification models and results from Chapter 4 provide the basis for Chapter 5 where the recent developments in the field of neural networks are used to provide model-induced and data-induced uncertainty maps besides the classification maps.

Finally in Chapter 6, the thesis is concluded with a summary of contributions and proposed future work.

Chapter 2

Background

2.1 Introduction

This chapter is devoted to reviewing the materials and methods required for the purpose of improving the quality of sea ice data assimilation states, with emphasis on regularized data assimilation and uncertainty estimation of sea ice retrievals from remote sensing data. First, a general overview of data assimilation and its techniques is given. Following that, the context of regularization for data assimilation as an inverse problem is reviewed. Then, sea ice parameters and methods to retrieve these parameters from remote sensing instruments are briefly reviewed to provide the prerequisites for the following content on ice water classification and uncertainty estimation.

2.2 Data Assimilation Overview

Data assimilation is a method for combining available observations, with a prior estimate of the current state of the system that is typically based on a numerical model, such as a forecast model, to find the best estimate of the current state of the system. The Regional Ice-Ocean Prediction System (RIOPS) developed by Environment and Climate Change Canada (ECCC), like any other operational ice prediction system, requires assimilation of different types of ice observations [89]. The reason for combining numerical model output with observational data, instead of relying solely on the observations is that in most cases, observations are sparse or partial in geophysics and they are also imperfect, and sometimes only indirectly measure the quantity of interest. Models are required to

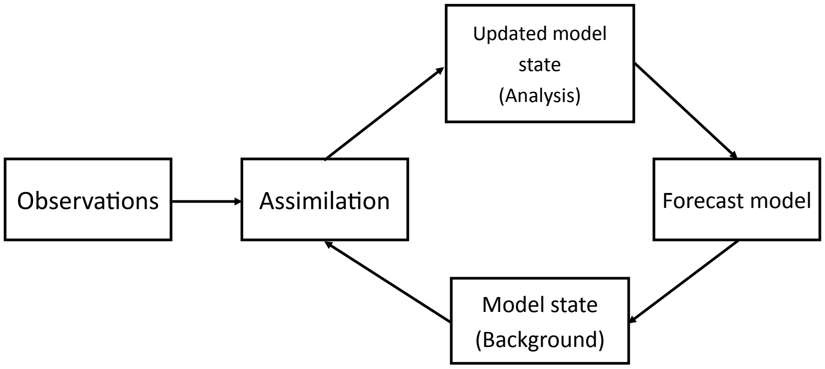


Figure 2.1: Forecast-assimilation cycle of data assimilation

interpolate available observations to unobserved regions or quantities. Since the beginning of data assimilation in 1940s, numerical weather prediction (NWP) is its most well-known application [25, 69]. Navigation systems [26], remote sensing [24] and pollution source estimation [109] are other applications of data assimilation.

2.2.1 Data Assimilation Cycle

Data assimilation, as described in Figure 2.1, is performed in cycles. Each cycle consists of two steps. In the first step, called the assimilation step, using the given model state (called the background state) and the observations, an optimal state estimate called the analysis is produced. In the second step, known as the forecast step, the analysis is given to the forecast model to be integrated forward in time. The outcome of this process is a new forecast or background state for the next cycle.

2.2.2 Notations and Operators

In this subsection, principal notation and operators of data assimilation are reviewed briefly.

- **State vector**: The first step in developing a data assimilation system for the desired application is mathematical formalization. This system needs a numerical representation of the system state, which is denoted as state vector $\mathbf{x}$. The best possible representation of reality as a state vector is called the true state $\mathbf{x}_t$. The prior estimate or initial guess of the state before the analysis is called the background, denoted as $\mathbf{x}_b$ and the optimal state estimate after the analysis is called the analysis, denoted as $\mathbf{x}_a$.

- **Observations**: The observational data values are gathered in observation vector $\mathbf{y}$. In practice, the size of the observation vector is usually different from the size of the state vector. The state vector has to be mapped from the state space to the observation space to be used in the analysis and to enable comparison between the state vector and the observation vector. The mapping function is called the observation operator denoted by H. This operator can be either linear or nonlinear. Also, in some cases the number of observations are fewer than the number of elements in the state vector [9].

- **Error variables**: Since the background and observation states are both not quite accurate, error models are used to reflect these inaccuracies.

  The error variables in a data assimilation analysis are:

  – Background errors: The difference between the true state and the background state is defined as

  $$\varepsilon_b = \mathbf{x}_b - \mathbf{x}_t, \tag{2.1}$$

  while the background error covariance matrix is

  $$\mathbf{B} = \overline{(\varepsilon_b - \bar{\varepsilon}_b)(\varepsilon_b - \bar{\varepsilon}_b)^T}. \tag{2.2}$$

  Here, $\bar{\varepsilon}_b$ describes the mean of the background error .

  – Observation errors: The difference between the true state and the observation state is defined as

  $$\varepsilon_o = \mathbf{y} - H(\mathbf{x}_t), \tag{2.3}$$

  with $\bar{\varepsilon}_o$ as the mean of the observation error, the observation error covariance matrix is

  $$\mathbf{R} = \overline{(\varepsilon_o - \bar{\varepsilon}_o)(\varepsilon_o - \bar{\varepsilon}_o)^T}. \tag{2.4}$$

  – analysis errors: The difference between the true state and the analysis state is defined as

  $$\varepsilon_a = \mathbf{x}_a - \mathbf{x}_t, \tag{2.5}$$

  with $\bar{\varepsilon}_a$ as the mean of the analysis error, the analysis error covariance matrix is

  $$\mathbf{A} = \overline{(\varepsilon_a - \bar{\varepsilon}_a)(\varepsilon_a - \bar{\varepsilon}_a)^T}. \tag{2.6}$$

The averages of errors are known as biases and their presence is the sign of a systematic problem in the assimilation system as they will produce bias in the analysis [27].

When the size of model state is n, the $\mathbf{B}$ matrix is a square and symmetric matrix of size $n \times n$. Similarly, the size of matrix $\mathbf{R}$ is $m \times m$ when observation vector has size m. The diagonal elements of these matrices contain error variances and the off-diagonal elements are cross-covariances between each pair of components in the state or observations for $\mathbf{B}$ and $\mathbf{R}$, respectively.

The error models will help the analysis to minimize its difference from the truth. Observation errors mainly include instrument errors and representativeness errors, which refer to the error of representing the model state on a discrete grid and the error in interpolating from the model grid to the observation locations. The error can also be due to the fact that the physical scales represented by the observations and the state vector may be different, which means they can capture different phenomena. For example, H can be a radiative transfer model transforming brightness temperature to surface temperature, etc. Background and analysis errors are related to the forecast error and these errors can be described by probability density functions (pdfs), typically Gaussian function is a popular one that can be fully characterised by a mean and covariance [38].

Table 2.1 briefly represents the essential parameters used in a data assimilation problem.

2.2.3 Data Assimilation Techniques

There are a variety of different techniques that have been developed to tackle the state estimation problem in the assimilation step of a data assimilation cycle. These techniques can generally be classified into two groups: sequential methods and variational methods. In sequential assimilation algorithms such as Kalman filter [67] or Best linear unbiased estimator (BLUE) [9], the system of equations needed for an optimal solution is solved explicitly while in variational algorithms like 3D-Var the equations are solved implicitly through the minimization of a cost function. Starting with the BLUE analysis, some popular assimilation techniques will be reviewed in this section.

Best linear unbiased estimator (BLUE)

Best linear unbiased estimator (BLUE) is an assimilation technique that tries to find the optimal analysis state under some basic assumptions [9]. The linear term in BLUE

Table 2.1: Parameters notation and description

notation	**description**	**size**
$\mathbf{x}_t$	true state	$n \times 1$
$\mathbf{x}_b$	background state	$n \times 1$
$\mathbf{y}$	observation vector	$p \times 1$
$\mathbf{x}_a$	analysis state	$n \times 1$
ε_b	background error	$n \times 1$
ε_o	observation error	$p \times 1$
ε_a	analysis error	$n \times 1$
H	observation operator	$n \rightarrow p$
$\mathbf{B}$	background error covariance matrix	$n \times n$
$\mathbf{R}$	observation error covariance matrix	$p \times p$
$\mathbf{A}$	analysis error covariance matrix	$n \times n$

means that we are optimizing a linear combination of the background and observation and unbiased term means that all the errors are assumed to be unbiased. Also, the best or optimal solution is calculated by minimizing the trace of the analysis error variance. In the following, all the assumptions are described in detail.

1. Linearized observation operator: For any $\mathbf{x}$ in the vicinity of the background state $\mathbf{x}_b$ we have
$$H(\mathbf{x}) - H(\mathbf{x}_b) \simeq \mathbf{H}[\mathbf{x} - \mathbf{x}_b] \tag{2.7}$$
where $\mathbf{H}$ is a linear operator. For the linear operator the above equation is obvious but for the nonlinear cases it could be calculated as
$$\mathbf{H} \equiv \frac{\partial H}{\partial \mathbf{x}} \in \mathbb{R}^{n \times p} \tag{2.8}$$
where p is the number of observations, and n is the length of the state vector $\mathbf{x}$.

2. Non-trivial errors: $\mathbf{B}$ and $\mathbf{R}$ are positive definite matrices.

3. Unbiased errors: $\overline{\varepsilon_b} = \overline{\varepsilon_o} = 0$

4. Uncorrelated errors: $\overline{\varepsilon_b \varepsilon_o^T} = 0$.

5. Linear analysis: our optimal state is a linear combination of the background state and observations.

6. Optimal analysis: the aim is to make the analysis state as close as possible to the true state (i.e. it is a minimum variance estimate).

Based on the fifth assumption, the best estimate has the form of a linear combination of the background estimate and the observation:

$$\mathbf{x}_a = \mathbf{W}\mathbf{x}_b + \mathbf{K}\mathbf{y} \tag{2.9}$$

where $\mathbf{W}$ and $\mathbf{K}$ are to be determined to make the analysis optimal. Substituting equations 2.1, 2.3 and 2.5 into 2.9 gives

$$\varepsilon_a + \mathbf{x}_t = \mathbf{W}\left(\varepsilon_b + \mathbf{x}_t\right) + \mathbf{K}\left(\varepsilon_o + H(\mathbf{x}_t)\right). \tag{2.10}$$

Recalling that the errors are unbiased and assuming the observation operator is linear, after applying the expectation operator to equation 2.10 we have

$$\mathbf{W} = \mathbf{I} - \mathbf{K}\mathbf{H}. \tag{2.11}$$

Now substituting equation 2.11 into equation 2.9 gives

$$\mathbf{x}_a = (\mathbf{I} - \mathbf{KH})\, \mathbf{x}_b + \mathbf{Ky} \tag{2.12}$$

and so

$$\mathbf{x}_a = \mathbf{x}_b + \mathbf{K}\,(\mathbf{y} - \mathbf{H}\mathbf{x}_b)\,. \tag{2.13}$$

The next step is to obtain **K**. Using equations 2.1,2.3,2.5 and 2.13 we have

$$\varepsilon_a = (\mathbf{I} - \mathbf{KH})\,\varepsilon_b + \mathbf{K}\varepsilon_o. \tag{2.14}$$

By developing the expression of $\varepsilon_a \varepsilon_a^T$ and taking its expectation, recalling the linearity of the expectation operator and the uncorrelated error assumption leads to

$$\overline{\varepsilon_a \varepsilon_a^T} = (\mathbf{I} - \mathbf{KH})\,\mathbf{B}\,(\mathbf{I} - \mathbf{KH})^T + \mathbf{KRK}^T. \tag{2.15}$$

The best unbiased estimate in BLUE analysis corresponds to the minimum analysis error variance. As the last step, the analysis error covariance matrix given by 2.15 is taken and its trace is minimized by setting its derivative with respect to **K** to zero. The minimization yields

$$\mathbf{K} = \mathbf{BH}^T(\mathbf{HBH}^T + \mathbf{R})^{-1} \tag{2.16}$$

and also

$$\mathbf{A} = (\mathbf{I} - \mathbf{KH})\,\mathbf{B}. \tag{2.17}$$

More detailed explanations on how to derive equations 2.16 and 2.17 exist in [9]. equations 2.13, 2.16 and 2.17 constitute the best linear unbiased estimator (BLUE) under the constraint of minimum variance. In these equations, **K** is called the gain, or weight matrix, and **A** is the analysis error covariance matrix.

Kalman filter

The BLUE analysis assumes that the observation and background error covariance matrices are fixed in time. However, this is a poor assumption when the background error changes with time while the model state evolves. The Kalman filter, introduced by Kalman [67] and Kalman and Bucy [68], is a method in which the background error covariance matrix also evolves in time in addition to the model state. The Kalman filter is a widely applied concept in time series analysis used in fields such as signal processing and econometrics [6]. The Kalman filter process is decomposed in two steps: analysis and forecast. The analysis step is the same as the BLUE analysis described in Section 2.2.3 except for the following notations and assumptions:

- variable i is added to include the time sequence. A model exists where $\boldsymbol{M}_{i \to i+1}$ indicates the model forecast operator from time i to $i+1$.

- model error $\eta(i)$ is defined as the deviation of the forecast prediction from the evolved true state, $\boldsymbol{M}_{i \to i+1}\mathbf{x}_t(i) - \mathbf{x}_t(i+1)$. The assumption is that the model is not biased and the model error covariance matrix denoted as $\mathbf{Q}(i)$ is known.

- analysis errors $\varepsilon_a(i) = \mathbf{x}_a(i) - \mathbf{x}_t(i)$ and models errors of the subsequent forecast $\boldsymbol{M}_{i \to i+1}\mathbf{x}_t(i) - \mathbf{x}_t(i+1)$ are mutually uncorrelated.

- forecast operator is linearized; i.e., for any $\mathbf{x}(i)$ in the close vicinity of $\mathbf{x}_a(i)$, we have $\boldsymbol{M}_{i \to i+1}[\mathbf{x}(i)] - \boldsymbol{M}_{i \to i+1}[\mathbf{x}_a(i)] = \mathbf{M}_{i \to i+1}[\mathbf{x}(i) - \mathbf{x}_a(i)]$.

Now the analysis step of the Kalman filter at each iteration performs a state analysis as

$$\mathbf{x}_a(i) = \mathbf{x}_f(i) + \mathbf{K}(i)[\mathbf{y}(i) - \mathbf{H}(i)\mathbf{x}_f(i)], \tag{2.18}$$

where $\mathbf{x}_f(i)$ is the forecast state vector and with the analysis error covariance

$$\mathbf{A}(i) = [\mathbf{I} - \mathbf{K}(i)\mathbf{H}(i)]\mathbf{B}(i), \tag{2.19}$$

and the Kalman gain computation as

$$\mathbf{K}(i) = \mathbf{B}(i)\mathbf{H}^T(i)[\mathbf{H}(i)\mathbf{B}(i)\mathbf{H}^T(i) + \mathbf{R}(i)]^{-1}. \tag{2.20}$$

In the forecast step the state forecast and the error covariance matrix are updated as follows

$$\mathbf{x}_f(i+1) = \boldsymbol{M}_{i \to i+1}\mathbf{x}_a(i), \tag{2.21}$$

and

$$\mathbf{B}(i+1) = \boldsymbol{M}_{i \to i+1}\mathbf{A}\boldsymbol{M}^T_{i \to i+1} + \mathbf{Q}(i). \tag{2.22}$$

Equation 2.22 is obtained by subtracting $\mathbf{x}_t(i+1)$ from equation 2.21 and using the linearity of the forecast operator:

$$\mathbf{x}_f(i+1) - \mathbf{x}_t(i+1) = \mathbf{M}_{i \to i+1}[\mathbf{x}_a(i) - \mathbf{x}_t(i)] + [\boldsymbol{M}_{i \to i+1}\mathbf{x}_t(i) - \mathbf{x}_t(i+1)], \tag{2.23}$$

which is equal to

$$\varepsilon_f(i+1) = \mathbf{M}_{i \to i+1}\varepsilon_a(i) + \eta(i), \tag{2.24}$$

where $\varepsilon_f(i+1) = \mathbf{x}_f(i+1) - \mathbf{x}_t(i+1)$ is the background error at $i+1$. Recalling that analysis error and models error are assumed to be uncorrelated, multiplying each side of the equation 2.24 by its transpose and taking the expectation yields equation 2.22. Even tough the basic Kalman filter is limited to a linear assumption, natural systems are often nonlinear. The nonlinearity can be associated either with the model or with the observation model or with both. The extended Kalman filter is an instance of the non-linear Kalman filter which tries to linearize the forecast model and observation operator but when the non-linearity is high, it can give particularly poor performance [64]. Related approaches include the ensemble Kalman filter and its variants [33, 34].

Variational data assimilation

The possible use of methods based on variational calculus in data assimilation was proposed by Sasaki [113, 114] in the late 1950s and 1960s. In contrast to BLUE analysis that explicitly solves the equations to find the best estimate, variational methods minimize a cost function to implicitly solve the problem. One such method is three-dimensional variational data assimilation (3D-Var) and an extension of this is four-dimensional data assimilation (4D-Var)[9].

The 3D-variational data assimilation scheme can be defined by a minimization problem with a cost function of the form

$$J(\mathbf{x}_0) = \overbrace{(\mathbf{x}_0 - \mathbf{x}_b)^T \mathbf{B}^{-1} (\mathbf{x}_0 - \mathbf{x}_b)}^{J_b} + \overbrace{(\mathbf{y} - H(\mathbf{x}_0))^T \mathbf{R}^{-1} (\mathbf{y} - H(\mathbf{x}_0))}^{J_o}. \tag{2.25}$$

This cost function tries to find a state, $\mathbf{x}_0$, that has the minimum distance from the background and observations where each distance is weighted by the error covariance matrix. The gradient of cost function J is

$$\nabla J = 2\mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}_b) - 2H^T \mathbf{R}^{-1}(y - H(\mathbf{x}_0)). \tag{2.26}$$

Lorenc showed if operator H is linear, the 3D-Var is strictly equivalent to the BLUE [95], since:

$$\begin{aligned}
\mathbf{x}_a = \operatorname*{argmin}_{\mathbf{x}_0} J(\mathbf{x}_0) \Rightarrow \nabla J(\mathbf{x}_a) = 0 \\
\Rightarrow 2\mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}_b) - 2\mathbf{H}^T \mathbf{R}^{-1}(y - \mathbf{H}\mathbf{x}_0) = 0 \\
\Rightarrow (\mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H})\mathbf{x}_0 = \mathbf{B}^{-1}\mathbf{x}_b + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{y} \\
\overset{\mathbf{x}=\mathbf{x}_a}{\Longrightarrow} \mathbf{x}_a = \mathbf{x}_b + (\mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{R}^{-1} (\mathbf{y} - \mathbf{H}\mathbf{x}_b)
\end{aligned} \tag{2.27}$$

and based on the Sherman-Morrison-Woodbury (SMW) formula

$$(\mathbf{B}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{R}^{-1} = \mathbf{B}\mathbf{H}^T(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1} = \mathbf{K}. \tag{2.28}$$

So, for 3D-Var for the case where the observation operator is linear, the analysis is given by,

$$\mathbf{x}_a = \mathbf{x}_b + \mathbf{K}\left(\mathbf{y} - \mathbf{H}\mathbf{x}_b\right) \tag{2.29}$$

which is equivalent to the BLUE analysis. Moreover, according to the set of equations in 2.27, the analytical solution for the 3D-Var minimization problem when the observation operator is linear can be represented as

$$\mathbf{x}_a = (\mathbf{B}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}(\mathbf{H}^T\mathbf{R}^{-1}\mathbf{y} + \mathbf{B}^{-1}\mathbf{x}_b). \tag{2.30}$$

In fact, when the number of elements in the state vector increases and/or the observation operator H is highly nonlinear, 3D-Var direct minimization may be less costly or more precise than taking the inverse of matrix $\mathbf{K}$. In contrast, when the size of the state vector is small, using the BLUE approach, which is a linear approximation, is more efficient than the evaluation of the cost function and its gradient at each iteration of the minimization process.

2.3 Regularization and Data Assimilation

This section aims to provide a general theoretical overview of data assimilation as an inverse problem and its relationship to regularization. Since the most popular regularization framework is Tikhonov regularization, this approach is reviewed here.

2.3.1 Data Assimilation as an Inverse Problem

Given the state, $\mathbf{x}_t$, the forward problem can be defined as that of determining the observations, $\mathbf{y}$, from this state having a definite physical or mathematical model H as

$$H(\mathbf{x}_t) = \mathbf{y}. \tag{2.31}$$

For example, if we observe temperature, currents and sea surface height at given points in space and time, observation operator H contains the dynamical model mapping the initial state of the ocean to these observations. Hence, data assimilation can be viewed as an inverse problem because the true state is unknown and the observations $\mathbf{y}$ are determined to try to recover $\mathbf{x}_t$. In addition, the fact that the number of observations is typically less than the number of elements in the state vector, coupled with errors in the observations and in the prior state estimate, as well as in the model, leads to a highly ill-posed inverse problem. So, in this case, instead of the true solution, the goal is to find a best solution by formulating the problem as a minimization problem. In this minimization problem, the true solution is approximated as the state that is closest to all observations, for example using a least-square approach:

$$\mathbf{x}_t = \arg\min_{\mathbf{x}} \|\mathbf{y} - H(\mathbf{x})\|_2. \tag{2.32}$$

However, since typically only part of the state vector is observed, even a least-square method will not lead to a unique solution and the problem is underdetermined. This problem is generally solved by introducing regularization terms that are added to the least-square penalty function to further constrain the solution.

2.3.2 Tikhonov (l_2-norm) Regularization

Recalling subsection 2.2.3, the 3D-Var data assimilation scheme is defined by a minimization problem to find optimal state $\mathbf{x}_0$ with a cost function of the form

$$J(\mathbf{x}_0) = \overbrace{(\mathbf{x}_0 - \mathbf{x}_b)^T \mathbf{B}^{-1} (\mathbf{x}_0 - \mathbf{x}_b)}^{J_b} + \overbrace{(\mathbf{y} - H(\mathbf{x}_0))^T \mathbf{R}^{-1} (\mathbf{y} - H(\mathbf{x}_0))}^{J_o}, \tag{2.33}$$

which can also be represented as,

$$J(\mathbf{x}_0) = \frac{1}{2}\|\mathbf{x}_0 - \mathbf{x}_b\|^2_{\mathbf{B}^{-1}} + \frac{1}{2}\|\mathbf{y} - H(\mathbf{x}_0)\|^2_{\mathbf{R}^{-1}}. \tag{2.34}$$

Under the assumption that the observation operator is linear ($H(\mathbf{x}_b) = \mathbf{H}\mathbf{x}_b$), the analytical solution of the minimization problem is

$$\mathbf{x}_a = (\mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H})^{-1} (\mathbf{H}^T \mathbf{R}^{-1} \mathbf{y} + \mathbf{B}^{-1} \mathbf{x}_b). \tag{2.35}$$

The error covariance matrices, $\mathbf{B}$ and $\mathbf{R}$, are positive definite so, the matrix ($\mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H}$) is always positive definite and as a result also has an inverse. However, since generally there are not enough observations available and the observations have errors, this problem might be very ill-conditioned. To study the stabilizing role of the background error covariance matrix, Johnson et al. [63] proposed to reformulate the classic variational data assimilation problem similar to the standard form of the Tikhonov regularization [126].

Accordingly, by letting $\mathbf{B} = \sigma_b^2 \mathbf{C_B}$ and $\mathbf{R} = \sigma_o^2 \mathbf{C_R}$ where $\mathbf{C_B}$ and $\mathbf{C_R}$ are the correlation matrices and using a change of variable $\mathbf{z} = \mathbf{C_B}^{-1/2}(\mathbf{x}_0 - \mathbf{x}_b)$, the cost function can be reformulated as

$$J(\mathbf{z}) = \|\mathbf{C_R}^{-1/2}(\mathbf{y} - \mathbf{H}\mathbf{x}_b) - \mathbf{C_R}^{-1/2}\mathbf{H}\mathbf{C_B}^{1/2}\mathbf{z}\|_2^2 + \mu^2\|\mathbf{z}\|_2^2, \tag{2.36}$$

where $\mu^2 = \sigma_o^2/\sigma_b^2$. If we define $\mathbf{f}$ and $\mathbf{G}$ as $\mathbf{f} = \mathbf{C_R}^{-1/2}(\mathbf{y} - \mathbf{H}\mathbf{x}_b)$ and $\mathbf{G} = \mathbf{C_R}^{-1/2}\mathbf{H}\mathbf{C_B}^{1/2}$, the cost function would be

$$J(\mathbf{z}) = \|\mathbf{f} - \mathbf{G}\mathbf{z}\|_2^2 + \mu^2\|\mathbf{z}\|_2^2. \tag{2.37}$$

Therefore, by solving

$$\mathbf{z}_a = \arg\min_{\mathbf{z}} J(\mathbf{z}), \tag{2.38}$$

the analysis can be obtained as

$$\mathbf{x}_a = \mathbf{x}_b + \mathbf{C_B}^{1/2}\mathbf{z}_a. \tag{2.39}$$

Having the above reformulated problem, Johnson provided new insights into the role of the background error covariance matrix on improving the condition number of the $(\mathbf{B}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})$ and thus stability of the classic variational data assimilation problem. Other regularization techniques can also be considered, which are discussed in Chapter 3.

2.4 Sea Ice Observations

2.4.1 Sea Ice Parameters

Operational ice monitoring requirements were one of the first applications of sea ice parameter retrieval from remote sensing data. These parameters mainly include ice concentration, ice extent, ice thickness, ice surface temperature, ice types and ice velocity.

The percentage of ice cover within a given area is defined as sea ice concentration. The considered area can be a footprint of a satellite sensor or a specified polygon covering a geographic region. Ice concentration is produced regularly on a daily or weekly basis in the form of maps or charts at a few centers such as Canadian Ice Service (CIS), EUMETSATs Ocean and Sea Ice Satellite Application Facility (OSISAF), NASA Goddard Space Flight Center (GSFC) and National Snow and Ice Data Center (U.S.A) (NSIDC). The operational ice charts of CIS, for instance, include estimates of the total ice concentration as well as

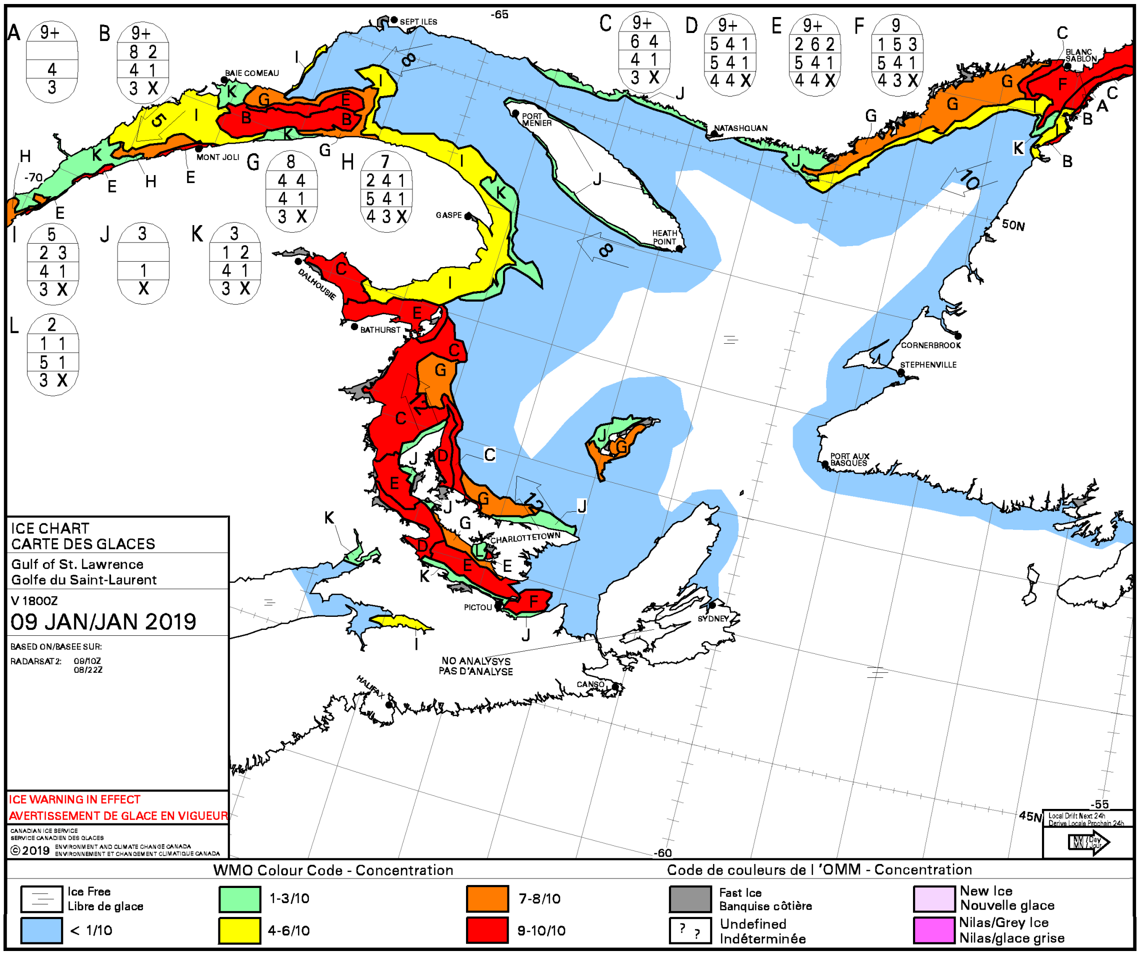


Figure 2.2: An example of daily ice chart for the Gulf of Saint Lawrence, January 9, 2019.

the partial concentration of presumed ice types. While ice concentration estimation is important for sea ice navigation and marine vessel and structure operations [48], it is also a key parameter in specifying the heat and moisture fluxes between the ocean and the atmosphere [116]. In Figure 2.2 a CIS daily ice chart is represented for the date of January 9, 2019 over the Gulf of Saint Lawrence. The ice chart decomposes the region into different polygons. Each polygon is associated with an egg code in an oval form that indicates the total concentration, partial concentration, stage of development and form of ice [97].

Using the concept of ice concentration, ice extent is defined as the area of sea where the ice concentration is beyond a specific threshold (usually 15%) [136]. Subsequently, ice edge is identified as the boundary of ice extent [22]. Ice extent is mostly used in monitoring the annual variation of the Arctic ice, which is an important indicator of global warming.

Sea ice thickness, as another parameter, is important for both climate studies and marine operations. While marine operators are more concerned with thick ice rather than thin ice due to navigational hazards, climate and weather models need more information on thin ice because of its essential impact on heat flux from the ocean to the atmosphere [118]. Information on sea ice temperature can also help indicate ice thickness when the ice is snow free and thin (less than 15cm) [66, 124].

Depending on the stage of development of sea ice, various ice types have been defined including new ice (weakly frozen crystals), nilas (<10cm thickness), young ice (10-30cm thickness), first-year ice (≥30cm thickness) and old ice (survived at least one summers melt) [97]. Ice type information is a vital part of ice chart products.

2.4.2 Sources of Observations

Remote sensing is a primary tool to monitor and provide information about sea ice parameters since the major volume of ice is located in remote areas at high latitudes, particularly in the polar regions. A great deal of remote sensing imagery used today in sea ice applications is provided by satellites. Different satellites and sensors have different unique characteristics, which makes them useful for specific applications. Here, a brief summary of some basic characteristics is provided.

Energy required for remote sensing of the earth is provided by electromagnetic radiation. The electromagnetic spectral regions used in remote sensing applications consist of optical, thermal infrared (IR), and microwave [53]. The optical region covers the visible (VIS), near IR (NIR) and shortwave IR (SWIR) bands. Optical sensors have limitation of sunlight and cloud-free sky availability. Sensors operating in middle-IR (MIR) and thermal infrared region (TIR) detect thermal emission. TIR sensors can discriminate between ice and water based on the temperature they capture from sea and ice surfaces under cloud-free sky. While observations in far-IR region (FIR) are not used in sea ice applications, microwave sensors in the millimeter and centimeter wavelength are the primary sensors for sea ice monitoring. The microwave region of the spectrum is quite large, relative to the visible and infrared, and there are several wavelength ranges or bands commonly used [131].

In the microwave region of the spectrum the surface signal is mostly unaffected by atmospheric conditions, especially for relatively large wavelengths (>10mm) [131]. Microwave radiation can penetrate through cloud cover, haze, dust, and all but the heaviest rainfall. So, microwave energy can be detected under almost all weather and environmental conditions and does not require sunlight. Passive microwave sensors measure emitted radiation while active sensors measure the backscatter of a transmitted signal after reflection or scattering off a surface. So far, passive microwave remote sensing has provided the most comprehensive long-term observations of sea ice data for global sea ice monitoring [118].

Passive microwave sensors detect the naturally emitted microwave energy within their field of view. They are typically radiometers or scanners and the amount of energy they detect depends on the temperature and moisture properties of the emitting object or surface. Earth surface radiation in the microwave range is very weak. Therefore, the fields of view must be large enough to detect the energy. Consequently, most passive microwave sensors are characterized by low spatial resolution (a few kilometers or tens of kilometers) and details of the sea ice, such as leads or other openings in the ice cover, are not easily detected, but in contrast the spatial coverage is large. Therefore, passive microwave observations are more suitable for large-scale or global monitoring. Passive microwave sensors are used mainly for ice concentration [21, 120, 127] and ice extent [16, 99] monitoring. In addition, estimation of a parameter such as ice thickness from passive microwave observations is a challenging task since the depth of the emitting layer, which depends on the wavelength and ice conditions, is typically limited to the few millimeters or centimeters. Therefore, observations are not usually useful for ice thickness estimation. For example, 1.4 GHz sensors onboard the soil moisture and ocean salinity (SMOS) satellite can measure ice thickness up to 50 cm [65, 66] and operational microwave frequencies in the range of 18-89 GHz are able to estimate ice thickness with an upper limit of 15-20 cm [102, 124].

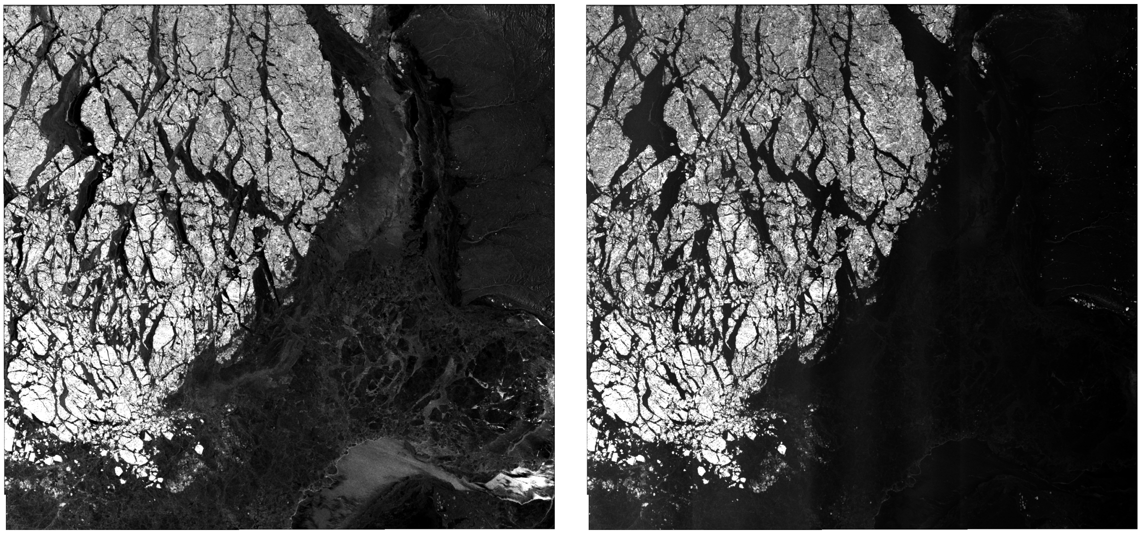
Active microwave sensors transmit a microwave signal towards the target and detect the backscattered portion of the signal. These sensors, also known as radars, are divided in two main categories: imaging and non-imaging. Altimeters [106] and scatterometers [93] are examples of non-imaging sensors. Compared to passive microwave sensor, for local observations such as determining the route for marine navigation through ice-covered water, imaging active microwave sensors are better applicable as they produce observations at finer spatial resolutions (for SAR around 100m or even higher) and also provide a relatively wide spatial coverage (even several hundreds of kilometers) [105].

The most commonly used active microwave sensor for operational ice monitoring in Canada is synthetic aperture radar (SAR). The first space-borne SAR, namely SEASAT, was launched in 1978 and ended the mission after 105 days. However, its 25m fine resolution imagery data were used to produce the first detailed sea ice motion maps and set the stage for more space-borne SAR missions. Identifying individual ice floes [54], producing an ice deformation grid [35], identification of openings in the ice cover and determining heat exchange from the ocean to the atmosphere [13] were successful application of the first SAR.

For the Canadian Ice Service (CIS), SAR data from the Canadian satellite RADARSAT-1 was the prime source of information for data analysis and ice charts products since 1996. RADARSAT-1 was replaced with RADARSAT-2 since its commissioning in 2008 (RADARSAT-2 was launched in December 2007 and RADARSAT-1 was decommissioned in March 2013). The most reliable current ice information is provided by SAR observations from the ongoing satellite missions such as Canadian RADARSAT-2, European Sentinel-1A and 1B, TanDEM-X, Japanese PALSAR-2 and Italian X-Band Cosmo-SkyMED.

One feature of the SAR sensors is their polarization which is described by two letters: the first one describes the emitted polarization and the second one describes the received polarization. Thereby, a SAR system using H and V linear polarizations can have HH, VV, HV and VH channels. RADARSAT-2, as the current main data source of CIS, carries an advanced SAR with different imaging modes and polarizations with a resolution from 3 to 100 metres. The HH images include more details than HV images but these are very sensitive to wind speed and incidence angle. In contrast, HV backscattering is not sensitive to the incidence angle and is less sensitive to wind speed but it suffers from banding effect that leads to distortion. At small incidence angles, wind roughened open water, new ice and first year ice are difficult to distinguish in HH images. However, HV images can provide information about the first year ice since water and new ice have zero or near zero backscatter in HV images. Figure 2.3 shows an example of HH and HV images of a dual-pol RADARSAT-2 image acquired on April 25, 2015 over the Beaufort Sea. Bright areas on the left side of the HH and HV images correspond to multi-year ice and the lighter grey region at the bottom right of the images is wind-roughened open water.

Even though SAR imagery is the primary source of data for generating sea ice charts in national ice centers for more than thirty years, a robust approach to retrieve ice concentration from SAR has not been developed [130]. Therefore, the required information is still extracted based on visual analysis of the image by trained ice analysts. The main challenge for developing an automated method for ice classification and concentration retrieval is the variation of backscatter signatures from different snow-covered sea ice and wind-roughened open water. SAR backscatter in ice-covered regions is influenced by many physical parameters, such as wind speed and direction, surface roughness, salinity, ice structure, melt, snow cover, temperature and floe distribution [14]. As an example, low backscatter of smooth ice surface overlaps with the low backscatter of calm open water. Similarly, high backscatter generated by wind driven ocean surface roughness overlaps with the high backscatter signature from rough young and first-year ice [118].



(a) HH

(b) HV

Figure 2.3: HH and HV bands of a dual-pol RADARSAT-2 image acquired on April 25, 2015 over Beaufort Sea. HH image includes more details but it is sensitive to wind speed and incidence angle.

2.4.3 Automated Ice/Water Classification of SAR Data

Given the difficulties discussed above about the complexity of discriminating ice from open water in SAR imagery of sea ice, here, some of the previous studies that have attemped to address this problem are now reviewed.

The overlapping characteristic of backscatter signals between ice and water resulted in the employment of image texture features such as gray level co-occurrence matrix (GLCM) to improve the quality of ice/water classifications. Holmes et al. introduced the use of texture features for sea ice classification of HV image of the Beaufort Sea [62]. They reported an overall accuracy of 65%. Clausi and Deng also used tone besides texture features in a study on pixel-based segmentation of RADARSAT1 images [19]. Their unsupervised approach used Markov random field (MRF) tuning on top of K-means clustering to classify both ice/water as well as ice types. The continuation of this study has led to an approach called Iterative Region Growing with Semantics (IRGS) which is an iterative region growing segmentation method followed by a support vector machine (SVM) classification to label the segments [133]. This method provides per-pixel ice type classification and is the first known successful end-to-end process for operational SAR sea ice image classification. The average reported classification accuracy of this method on 20 RADARSAT-2 scenes over the Beaufort Sea with respect to manually drawn ice charts is 96.5 % [20, 87, 104].

SVM classification has also been employed by Liu et al. to classify RADARSAT-2 images using backscatter and GLCM texture features as well as ice concentration [92]. In this study, first ice concentration is trained by SVM using 10 backscatter and texture bands as input. Then, another SVM is implemented on the retrieved ice concentrations besides the other 10 bands and finally the results are labeled into multiple sea ice types by applying a decision tree. Zakhvatkina et al. also applied SVM for the problem of ice/water classification using 24 RADARSAT-2 training images with backscatter and texture features [134]. Their algorithm was applied on a total of 2705 RADARSAT-2 scenes and validated against manually derived ice charts of the Norwegian Meteorological Institute. The reported accuracy was around 91%.

Thresholding is another classification approach that has been investigated in a number of studies. In a basic version, pixels are classified into two classes based on their grey level value using a cut-off threshold [15]. This cut-off threshold can be defined globally for the whole image or locally for each region. Dynamic thresholding has been introduced by Haverkamp et al. [58] and further continued by Soh and Tsatsoulis [119]. Karvonen et al. developed an algorithm for ice-water discrimination also using RADARSAT-1 images [72]. In this approach the images are first segmented based on their intensities and then segmentation is further refined by another segmentation based on local autocorrelation. After this, thresholding followed by filtering is applied to label the segments. The study reported 90% accuracy compared to ice charts for the Baltic Sea. This approach has been used by the Finnish Meteorological Institute (FMI) [70]. In another study, Karvonen has compared the segment-wise autocorrelation algorithm with another algorithm based on segment-wise edge information using both Envisat ASAR and RADARSAT-2 data [71].

Berg and Eriksson investigated the use of neural networks for the classification of ice and water as the first step of their ice concentration retrieval problem [8]. In this study, image backscatter autocorrelation is utilized as neural network input for training against 41 sea ice charts over the Baltic Sea. The results reported accuracy of 94% over open water and 87% for sea ice class with root mean square error of 6.7% in the sea ice concentration estimation. In another study, Zakhvatkina et al. extracted 9 GLCM texture features to train a neural network classifier which was then tested on 20 ENVISAT SAR images, resulting in average classification accuracy about 80% [135]. Ressel et al. also utilized neural networks on GLCM features of TerraSAR-X ScanSAR data acquired in spring 2013 over Western Barents Sea [108]. Their obtained average accuracy was above 70%. In a recent study, Aldenhoff compared ice/water maps generated from Sentinel-1 and

ALOS-2 PALSAR-2 SAR images over Fram Strait [2]. Their classification approach which uses backscatter intensities, the incidence angle and the autocorrelation to train a neural network, achieved an overall accuracy about 86%.

In a recent study, Komarov and Buehner [76] used a logistic regression method to model the probability of the presence of ice within areas of 2.05km×2.05km. For that purpose, a relatively large dataset containing thousands of RADARSAT-2 ScanSAR dual-polarization images along with their corresponding Canadian Ice Service (CIS) image analysis for verification, was collected. That dataset covers all seasons and all Canadian and adjacent Arctic regions that are monitored by CIS from 2010 to 2016. This ice/water retrieval approach, which only labels samples with high level of confidence, was claimed to be suitable for assimilation into Regional Ice-Ocean Prediction System (RIOPS). This study has been further continued by the authors to employ adaptive thresholding [77] and ice motion [78] in the classification procedure.

2.4.4 Observation Uncertainty Estimation in Sea Ice Retrievals

Use of retrieved sea ice parameters in downstream applications such as forecasting or climate studies is hindered due to the lack of uncertainty information provided with the retrievals. Here, some research with the goal of estimating uncertainty of sea ice parameters inside the retrieval approach are reviewed.

In a project conducted by EUMETSATs ocean and sea ice satellite application facility, a sea ice climate record of sea ice area and extent covering the period from 1978 to 2015 has been generated based on SMMR, SSM/I, and SSMIS measurements [127]. Besides the ice concentration algorithm used for sea ice extent calculation, a new sea ice concentration uncertainty estimation algorithm has also been developed. The estimated uncertainties are defined as the combination of two independent uncertainty components. The first is the tie-point uncertainty, which is derived from measurements and representativeness error. The second is the uncertainty caused by employing coarse resolution passive microwave measurements (footprint of 30-70 km) in finer grid spacing (typically 12.5 or 25 km). The results have shown that for open water and 100% ice covered areas this component of uncertainty is zero and we have the minimum total uncertainty (around 5%), in other cases it is the dominating component of the total uncertainty (around 12%).

The benefit of accounting these provided uncertainties to improve summer ice concentration and ice thickness forecasts in a data assimilation experiment are shown in a study by Yang et al. [132]. It was found that imposing the estimated uncertainties will improve the accuracy of ice concentration forecasts.

In a separate study, the uncertainty of ice concentrations retrieved by the NT2 algorithm is estimated [10]. NT2 uses passive microwave brightness temperature as well as simulated microwave brightness temperatures as input to estimate sea ice concentration. The uncertainty in this study is defined as the standard deviation of the ice concentrations having the 20 smallest δR values. In this definition, δR is the sum of differences between the observed and simulated inputs to the sea ice concentration retrieval algorithm. The experiments on regional and seasonal variations has found that ice concentration uncertainties are higher in areas with new ice and deep snow and at the onset of melting and during the melting season.

In a study by Zygmuntowska et al. Monte Carlo approach is used to calculate the uncertainty of sea ice volume and thickness which are defined as a function of sea ice area, density and snow depth [137]. The Monte Carlo approach repeatedly calculates the results by random sampling from the PDF of each of the input parameters and reporting the variance of results as the uncertainty. The total uncertainty is derived by simultaneously iterating through PDF of all the three parameters while the uncertainty of a single parameter is calculated by iterating through PDF of that single parameter and fixing the other two parameters at the mean of their distributions. The ice thickness experiments based on ICEsat freeboard measurements of 2005-2007 revealed higher uncertainty in period of October/November rather than February/March. Moreover, snow depth, with up to 70% contribution, was reported as the main contributor to the total uncertainty and ice density has about 30-35% contribution. Ice density contribution in the ice thickness uncertainty was higher in October/November because of small snow cover in that period of year and sea ice area contribution in total uncertainty was always below 10%. Ice volume total uncertainty was observed to follow almost similar pattern of contributions from each parameter.

All the methods described above, provide uncertainties with their products. However, there is still no method to provide uncertainty with ice/water labels or SAR retrieved observations. Neural network-based ice/water classification approaches suffer from lack of uncertainty in their products. The problem of uncertainty estimation for a SAR-based ice/water classification problem is investigated in Chapter 5 using the recent advances in the neural network studies.

2.5 Summary

This chapter provided the required background information for investigation of assimilating using regularization framework and uncertainty estimation of sea ice observations. The chapter started with reviewing basics of data assimilation and some popular data assimilation techniques, particularly the 3D-Var method. Moreover, the context of regularization, specifically Thikhonov regularization, was described in the field of data assimilation. After briefly discussing sea ice parameters and remote sensing sources of observations, specifically microwave remote sensing, some popular approaches on the classification of ice and open water based on SAR data were reviewed. The chapter ended by reviewing previous studies that addressed the problem of sea ice uncertainty estimation in the retrieval process.

Chapter 3

Data Fusion and Data Assimilation of Ice Thickness Observations

3.1 Introduction

Assimilation of sea ice data in case of noisy observations while the state vector exhibits sparsity (a large fraction of elements that are close to zero) in the real or transformed domain, can be improved by using sparse regularization. In this chapter, first, evidence of sparsity in the ice thickness derivative domain is demonstrated using both airborne survey data, and submarine sonar data. Next, the impact of using a sparse variational regularization framework, is demonstrated using both data fusion experiments and data assimilation experiments using a 1-D sea ice model.

3.2 Sparsity of Ice Thickness Observations

At the small scales, similar to observations of precipitation and other geophysical variables [32, 37], sea ice no longer behaves as a continuum [36], and exhibits sharp features or discontinuities. To visualise sharp features in the sea ice state, ice thickness measurements obtained using an airborne electromagnetic (AEM) sensor during an aircraft survey over the Beaufort Sea are examined. The AEM sensor measures the ice+snow thickness, which we will refer to as the ice thickness [51], with a spatial distance of 4.5m to 6m (ground distance) between measurement points. The data used here were acquired on April 20 2015 in the Beaufort Sea. The measurement sample points indicating the flight path are shown in Figure 3.1a. For more information about AEM ice thickness measurements see [52].

In the Beaufort Sea, the ice cover is highly heterogeneous, and consists of a mixture of thin and thick ice, first-year and multi-year ice, with many ridges. The ridges are spatially localized regions with higher ice thickness than the surrounding ice. While the AEM measurement is known to underestimate the maximum thickness of pressure ridges, it has high accuracy over level ice ($\pm 0.1m$) [50, 52] and it can represent the variability in ice thickness at small spatial scales, as can be seen in Figure 3.1b.

A state of interest is sparse in a real or transformed domain if its number of zero elements in that domain is significantly large compared to its dimension in the state space. To demonstrate the sparsity in the derivative field of AEM ice thickness, various methods are used. In a sparse distribution, the distribution is more peaked at zero, indicating many points with values close to zero, with fatter tails, as compared to a Gaussian distribution. The histogram of the first order spatial derivative of the AEM ice thickness measurements (taking into account all data points in Figure 3.1b) is shown in Figure 3.2a. Overlaid on this distribution are fitted Gaussian, Laplacian and generalized Gaussian distributions. The generalized Gaussian distribution is defined as

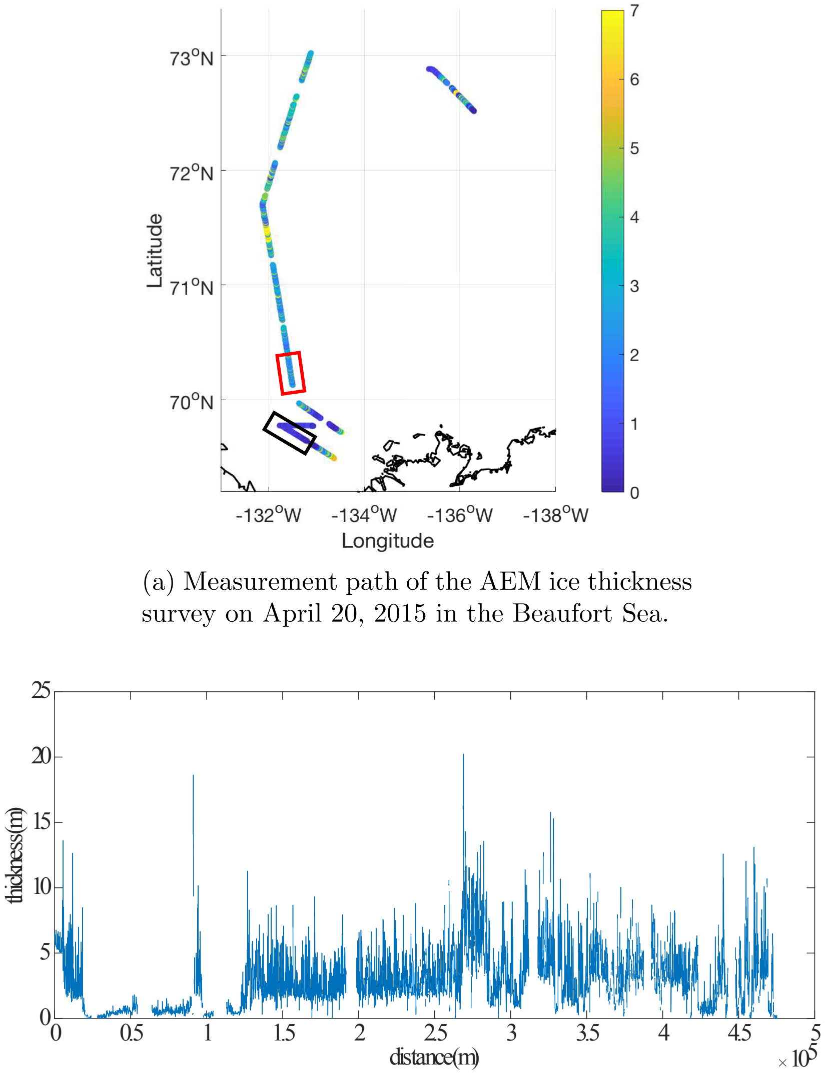
$$P_X(x) = \frac{p}{2\sigma\Gamma(1/p)} \exp\left(-|\frac{x}{\sigma}|^p\right), \tag{3.1}$$

where $\Gamma(z)$ is the Gamma function and p and σ are non-negative parameters describing the shape and width of the density, respectively. For special cases of $p = 1$ and $p = 2$, this probability distribution corresponds to Laplacian and Gaussian distribution respectively. The number of bins in the histograms of Figure 3.2 was set to the maximum of the Sturges and Freedman Diaconis estimators [39, 123] and the AEM measurement points were resampled to have an equal spacing of 7m as this simplifies application of the methods used.

Figure 3.2a shows that the data exhibits a symmetric representation in the derivative space with a large mass around zero. The large number of zero coefficients in the derivative domain is caused by the uniformity of ice thickness over a large area, whereas the large tails are associated with the ridges and/or openings in the ice cover. Note the density of the histogram values around zero is greater than the fitted Gaussian distribution.

For comparison, histograms of the first order spatial derivatives of sea ice draft from submarine upward looking sonar [103] and ice thickness retrieved from Cryosat data [1] [85] are also shown in Figure 3.2. Submarine data in Figure 3.2b are from an archive of data

[1] http://www.cpom.ucl.ac.uk/csopr/seaice.html



(a) Measurement path of the AEM ice thickness survey on April 20, 2015 in the Beaufort Sea.

(b) The sequence of the AEM thickness data for the measurement path shown in panel (a)

Figure 3.1: Sea ice thickness measurements from airborne electromagnetic (AEM) sensor in the Beaufort Sea acquired on April 20th 2015. (a) Sea ice thickness [m] is represented by the colorbar. The red rectangle outlines a region with deformed first year ice (FYI) while the black rectangle outlines a region with thinner and smoother first-year ice. These regions were determined from visual analysis of a SAR image acquired on April 19, 2015. (b) Sequential representation of the AEM data shown in panel (a).

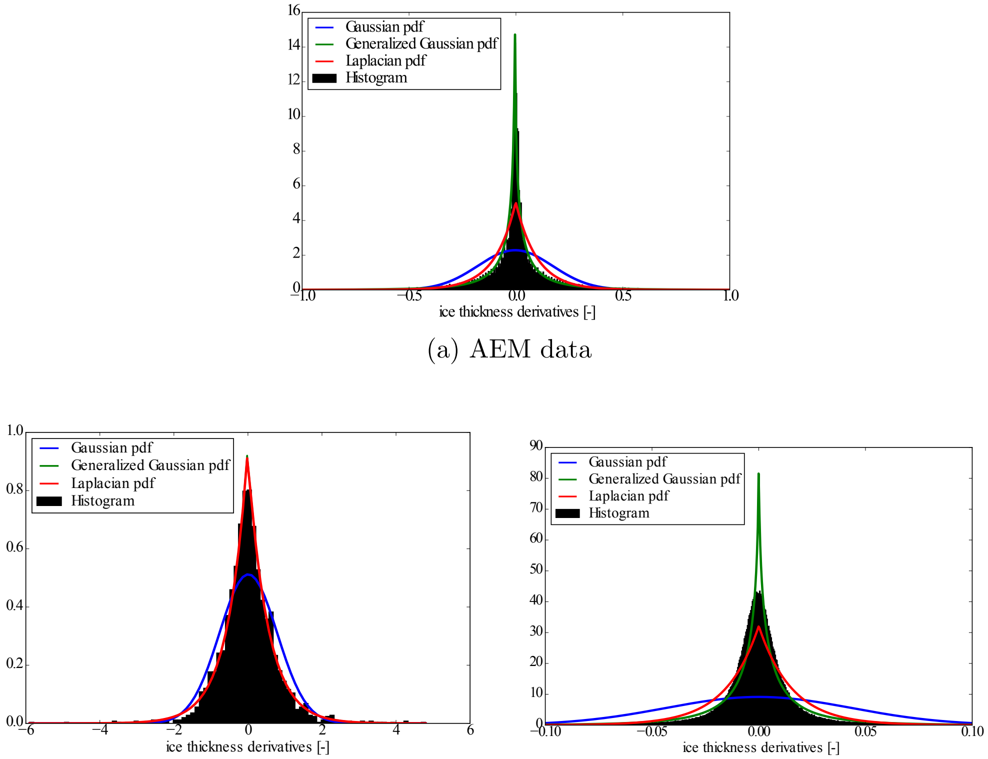
acquired in Arctic ocean in July 2005 and is resampled at 10m. It was expected that these data would be less sparse than the AEM data, since the underside of the ice is known to be irregular, containing fewer regions than the surface that can be considered level [110]. This is supported by the histogram in Figure 3.2b, where it shows that while the fit is closest to the Laplacian distribution, derivative coefficients have lower concentration around zero than that of the AEM data showing that AEM data peak is sharper. As another example of ice thickness sparsity, Figure 3.2c shows the histogram of the derivative field of the Cryosat sea ice thickness measurements at 5km grid from March and April of 2015. The distribution of ice thickness derivatives is highly peaked around zero, showing a better fit to Laplacian and Generalized Gaussian distributions rather than a Gaussian distribution.

To quantitatively describe the goodness of fit of the represented distributions in Figure 3.2 to the ice thickness derivative field distribution, Kullback-Leibler divergence (D_{KL}) [82] is used. D_{KL} is a measure of the distance between two distributions. For the discrete probability distributions P and Q it is defined as

$$D_{KL}(P\|Q) = \sum_{i} P(i) \log \frac{P(i)}{Q(i)} \tag{3.2}$$

where i is the size of the discrete probability distributions defined by the number of bins in the histogram. Table 3.1 represents the D_{KL} values for the three datasets in Figure 3.2. For AEM data, the D_{KL} values of two subsets of data, selected as smooth and deformed FYI corresponding to the black and red boxes in Figure 3.1 respectively, are also included. For each distribution, D_{KL} was calculated by substituting the histogram bin values for P and the fitted distribution values for each bin by Q into equation (3.2). The D_{KL} values of the first column (AEM_all) reflect what can be seen in the histogram, which is that the distribution of the AEM ice thickness derivatives is statistically closer to a Laplacian or generalized Gaussian distribution (with $p = 0.8$ in equation (3.1)) than a Gaussian distribution. While the D_{KL} values for the Cryosat data are relatively large, the Laplacian distribution is a better fit in terms of D_{KL} values to the AEM data and other datasets compared to Gaussian distribution.

To further evaluate the fit of the proposed distributions to the AEM data, a Q-Q plot [18] is used. In a Q-Q plot the sample data points are plotted against the quantiles of an assumed theoretical distribution in such a manner that the points should form approximately a straight line. Departure from the straight line indicates departure from the specified distribution. Figure 3.3 shows that Laplacian and generalized Gaussian distribution are better fits to the data in comparison with a Gaussian distribution.



(a) AEM data

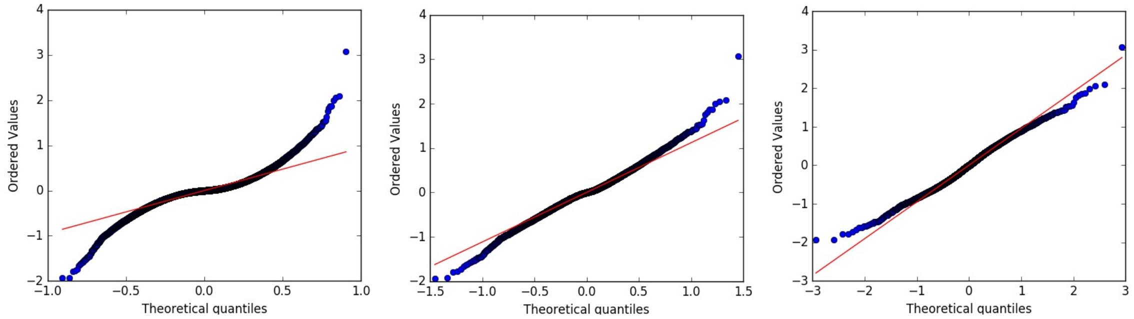
(b) Submarine data

(c) Cryosat data

Figure 3.2: Normalized histograms of the ice thickness spatial derivative fields of AEM data (a), submarine data (b), and Cryosat data (c) with the fitted Gaussian, Laplacian and generalized Gaussian distributions. All histograms show more similarity to generalized Gaussian and Laplacian distributions rather than a Gaussian distribution. For the submarine data (b), the fitted generalized Gaussian distribution and the Laplacian distribution overlap.

Table 3.1: Kullback-Leibler divergence (D_{KL}) between different ice thickness datasets and their fitted distributions on the derivative field. A lower number indicates a closer fit to the specified distribution. AEM_all refers to all AEM thickness measurements shown in Figure 3.1, while AEM_smooth_FYI and AEM_deformed_FYI refer to the data in the black and red boxes in Figure 3.1 respectively.

Distribution	AEM_all	AEM_smooth_FYI	AEM_deformed_FYI	Cryosat	Submarine
Gaussian	0.44	0.47	0.3	0.81	0.13
Laplacian	0.15	0.12	0.07	0.23	0.03
Generalized Gaussian	0.02	0.08	0.04	0.06	0.03



(a) Gaussian distribution

(b) Laplacian distribution

(c) Generalized Gaussian distribution

Figure 3.3: Q-Q plot of the data derivative field against Gaussian, Laplacian and generalized Gaussian distributions. The horizontal axes are the quantiles of the fitted distributions and the vertical axes are the ordered values of the derivative field of AEM data. Departure from the straight line indicates departure from the specified distribution.

3.3 Data Assimilation Method

The data assimilation method used here is a regularization form of 3D-var data assimilation. In image processing it has been shown that l_1-norm regularization gives a better performance when sharp edges need to be recovered [56]. The reason for the edge preserving property of the l_1-norm is that since edges in images lead to outliers in the regularization term, unlike l_2-norm, l_1-norm does not penalize the edges. In addition, it has been shown that with very high probability, the l_1-norm regularization promotes sparsity in the solution [30].

3.3.1 Application of l_1-norm in Data Assimilation

The applicability of employing l_1-norm regularization in a real data assimilation experiment has been investigated by Ebtehaj et al. [31]. The motivation of their work was to improvement the quality of rainfall estimation when the data comes with small-scale high-intensity extreme features in case of having coarse resolution sensor observations. By using Huber regularization in the derivative space of a precipitation data, the achieved results were improved compared to the ones when the Tikhonov regularization is used. They used the following cost function:

$$J(\mathbf{x}_0) = \frac{1}{2}\|\mathbf{x}_0 - \mathbf{x}_b\|^2_{\mathbf{B}^{-1}} + \frac{1}{2}\|\mathbf{y} - H(\mathbf{x}_0)\|^2_{\mathbf{R}^{-1}} + \lambda\|D\mathbf{x_0}\|_{\text{Hub}}, \tag{3.3}$$

with λ as the regularization factor. The Huber norm is also defined as

$$\|\mathbf{x}\|_{\text{Hub}} = \sum_{i=1}^{n} \rho_T(\mathbf{x}_i) \tag{3.4}$$

where,

$$\rho_T(\mathbf{x}) = \begin{cases} \mathbf{x}^2, & |\mathbf{x}| \leq \tau \\ \tau(2|\mathbf{x}| - \tau), & |\mathbf{x}| > \tau \end{cases}. \tag{3.5}$$

The Huber norm is a hybrid norm with similar behaviour to the l_1-norm for values greater than the τ and similar to the l_2-norm for smaller values. Therefore, Huber regularization is a smooth approximation of the l_1 regularization.

The reason for using this type of norm was the claim that the choice of regularization type in a regularization framework depends on prior knowledge about the underlying state of interest. Ebtehaj reported that using the regularization term in an assimilation problem is statistically equivalent to the maximum a posteriori (MAP) estimator. Consequently, using an l_2-norm means assuming a Gaussian distribution as the prior probability model of the desired state while using an l_1-norm term means assuming a Laplacian distribution. In a following test, the histogram of the rainfall derivatives with the fitted Gaussian and Laplacian distribution was plotted where the shape of the histogram was in close agreement with the fitted Laplacian distribution.

3.3.2 Mixed l_1-l_2-norm Regularization

In Section 2.3.2 it was shown that the cost function of the classic 3D-variational data assimilation problem can be reformulated to the standard form of Tikhonov (l_2-norm) regularization as

$$J(\mathbf{z}) = \|\mathbf{C_R}^{-1/2}(\mathbf{y} - H\mathbf{x}_b) - \mathbf{C_R}^{-1/2}\mathbf{H}\mathbf{C_B}^{1/2}\mathbf{z}\|_2^2 + \mu^2\|\mathbf{z}\|_2^2, \tag{3.6}$$

Freitag et al. [40] proposed their mixed regularization framework with a cost function as

$$J(\mathbf{z}) = \|\mathbf{C_R}^{-1/2}(\mathbf{y} - H\mathbf{x}_b) - \mathbf{C_R}^{-1/2}\mathbf{H}\mathbf{C_B}^{1/2}\mathbf{z}\|_2^2 + \mu^2\|\mathbf{z}\|_2^2 + \delta\|D\mathbf{x}_0\|_1. \tag{3.7}$$

where $\mathbf{x}_0 = \mathbf{x}_b + \mathbf{C_B}^{1/2}\mathbf{z}$, $\|.\|_1$ indicates l_1-norm (Laplacian distribution), and $\delta > 0$ is another regularization parameter and has to be chosen empirically. As δ increases, the gradient of the solution will be more sparse. Details about how to set this parameter are discussed in Section 3.4.3. In equation (3.7) the first order derivative can be approximated by applying matrix $D \in \mathbb{R}^{(n-1)\times n}$ to the state vector, where $\mathbf{D}$ can be written as

$$\mathbf{D} = \begin{bmatrix} -1 & 1 & 0 & ... & \\ 0 & -1 & 1 & 0 & ... \\ & \ddots & \ddots & \ddots & \\ & \cdots & 0 & -1 & 1 \end{bmatrix}_{(n-1)\times n} . \tag{3.8}$$

This operator corresponds to a first-order differencing, with equal spacing between data points. Other approximations to the first order derivative operator were tested, but the results were not found to be sensitive to this choice. This representation in equation 3.8 is chosen for simplicity.

As discussed in Section 3.2 the ice thickness observations exhibit sparsity in the derivative domain and have a closer fit to both the Laplacian and generalized Gaussian distribution as compared with the Gaussian distribution. Even though the generalized Gaussian distribution with $p = 0.8$ is the best fit for these observations, assuming this distribution for the regularization results in a non-convex optimisation, which might not converge easily. Therefore, assuming the Laplacian distribution as the prior model for the gradient field is the approach taken in this study.

3.4 Experimental Method

To evaluate the advantage of using the mixed regularization framework, two different regularized variational data assimilation systems are evaluated. The first system, l_2 regularization, minimises the objective function given by equation (3.6), while the second system, l_1-l_2 regularization, minimises the objective function described by equation (3.7). Note that the l_1-l_2 regularization method when $\delta = 0$ is equivalent to the l_2 method. The minimisation problems are reformulated to a quadratic programming using an approach introduced by [41] and solved by Python CVXOPT package [4]. These two systems were analysed by first carrying out a set of data fusion experiments, where the analysis is not cycled through a model. The objective of the first set of experiments is to look at the impact of the mixed regularization on the analyses, without the complications introduced by model physics and resampling of the AEM data to the model grid. The data fusion experiments are followed by a set of data assimilation experiments where a toy (1-D) sea ice model is used (Appendix B).

For both the data fusion and data assimilation experiments the background error covariance matrix is defined as

$$\mathbf{B} = \Sigma \mathbf{C_B} \Sigma \tag{3.9}$$

where Σ is a diagonal matrix with diagonal elements that are the background error standard deviation, σ_b, and $\mathbf{C_B}$ is the background error correlation matrix. The error correlation function is calculated based on the compactly supported fifth-order piecewise polynomial correlation function proposed by [45]. Note that in this formulation the error correlation function is an explicit function of the specified length scale and the spacing between points. The observation error covariance matrix, $\mathbf{R}$, is defined in the same way as the $\mathbf{B}$ matrix using error standard deviation σ_o and the same correlation function. Both the l_1 and l_1-l_2 systems require calculation of the square root of the background error correlation matrix. This is a computationally expensive operation, that scales as $O(n^3)$, where n is the dimension of the problem. For the present study, n is small ($n \approx 400$) and the square root is computed using Cholesky decomposition, which takes about 5% of the total time of an analysis cycle. For a larger, operational sea-ice data assimilation system, a method to compute the background error correlation matrix square root using a diffusion equation

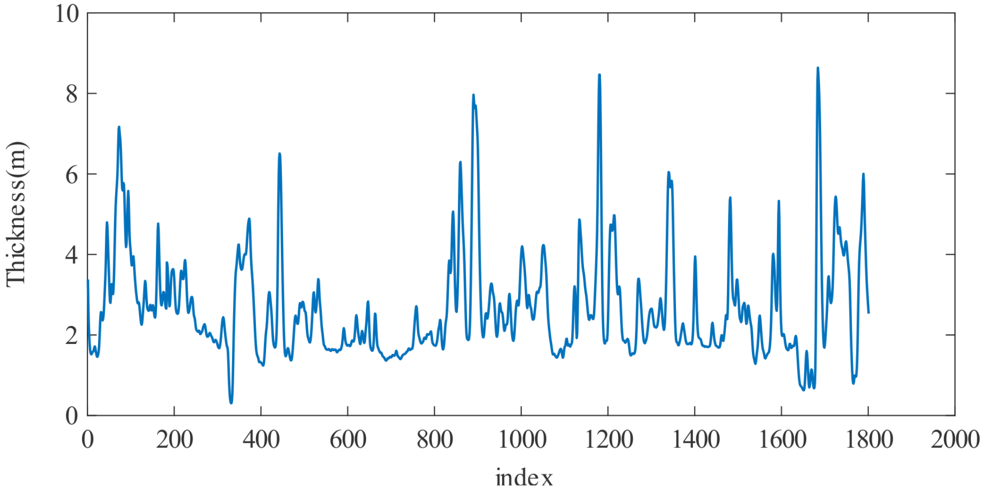


Figure 3.4: The deformed FYI sequence of the AEM thickness data outlined in Figure 3.1 with red rectangle. The data points are spaced at 7m.

is described in [17]. A similar method could be used for the l_1-l_2 method described here. An additional, potentially expensive, part of the algorithm is the inversion of the square root of the observation error correlation matrix. Large scale systems typically assume a diagonal observation error correlation matrix, which makes this operation trivial. Methods for large scale systems utilising non-diagonal observation error correlation matrices are still a topic of research in the operational data assimilation community.

The accuracy of the analyses is evaluated using six measures: (1) mean absolute error (MAE), (2) root mean squared error (RMSE), (3) kurtosis, $\text{KURT}[\mathbf{x_a}] = \frac{E\left[(\mathbf{x}-E[\mathbf{x}])^4\right]}{(E[(\mathbf{x}-E[\mathbf{x}])^2])^2}$, where $E[\cdot]$ represents the expectation operator, (4) MAE of the derivative field (DIFF_MAE) (5) RMSE of the derivative field (DIFF_RMSE), and (6) kurtosis of the derivative field ($\text{KURT}[D\mathbf{x_a}]$). Kurtosis of the state is used to show tail of the state's probability distribution so that, an analysis with high kurtosis has retained more sharp features. Reported results are averaged over 40 simulations where the additive random noise for background and observation states are different for each simulation. Note that the background and observation error standard deviation ratio (μ) was held constant at $\mu = 1$ for all experiments. Details specific to the data fusion and data assimilation experiments are now given.

3.4.1 Data Fusion Experiments

The deformed FYI subset of the AEM data shown in Figure 3.1 is selected as the true state ($\mathbf{x}_t \in \mathbb{R}^n$ where $n = 1800$). This sample of the data, shown in Figure 3.4, was found to include more sharp features as compared to the smooth FYI sample. The background ($\mathbf{x}_b \in \mathbb{R}^n$) and observations ($\mathbf{y} \in \mathbb{R}^n$), consist of the true state with additive random noise sampled from a Gaussian distribution with mean zero and covariance matrix $\mathbf{B}$ and $\mathbf{R}$ respectively. For both background and observation error covariance matrices, experiments were carried out over a range of correlation length scales, measured in meters, corresponding to $L = \{0, 20, 50, 100, 500\}$ and a range of δ values. It was found the values of the error measures (given below) increased with increasing δ for $\delta > 1$. Hence the performance of the l_1-l_2-norm regularization method is reported for a range of δ values between 0 and 1.

3.4.2 Data Assimilation Experiments

For the data assimilation experiments the initial true state was based on the entire set of AEM ice thickness measurements acquired on April 20th, shown in Figure 3.1b. The data points were averaged over segments of 1km in length. This was done to obtain a state at a spatial resolution closer to that of interest in operational sea ice forecasting [7, 12, 100], in addition to one where the continuum assumption used in the ice model should be valid [36]. When this true state was used to initialise the sea ice model it was found the ice did not move very significantly over the 72h period. This was likely because the ice concentration was close to 100% over the entire domain and the model was run with thermodynamics off to simplify analysis of the results. This means the ice concentration can only change by openings in the ice cover, which may be difficult to obtain due to the high ice pressure when the ice concentration is close to 100% and the lack of solid boundaries in the domain. To create open water regions in the ice cover the thickness everywhere was reduced by 0.5m and the points with negative thickness were set to zero. This state was used to initialise the ice model which was run forward in time to provide true states at $t > 0$, referred to in this paper as the true model run (Appendix B). While this state differs from the actual truth in part due to model error, such an approach should be sufficient for evaluating the difference between l_2 and l_1-l_2 regularization methods. Each of the experiments (truth, l_1-l_2 and l_2) was run with identical atmosphere and ocean states, which provide the forcing to the sea ice model.

The data assimilation experiments were initialized with a background state consisting of $\mathbf{x}_b = \mathbf{x}_t + \epsilon_b$ where ϵ_b is a sample generated from the $\mathbf{B}$ matrix described in the previous section. Every 6h the model state was updated by assimilating a set of observations generated by perturbing the current state of the true model run at that time according to $\mathbf{y} = \mathbf{x}_t + \epsilon_o$ where ϵ_o is a sample generated from the $\mathbf{R}$ matrix described in Section 4. Data assimilation experiments were carried out for background and observation correlation length scales corresponding to 0 and 10km.

3.4.3 Choice of Regularization Parameter

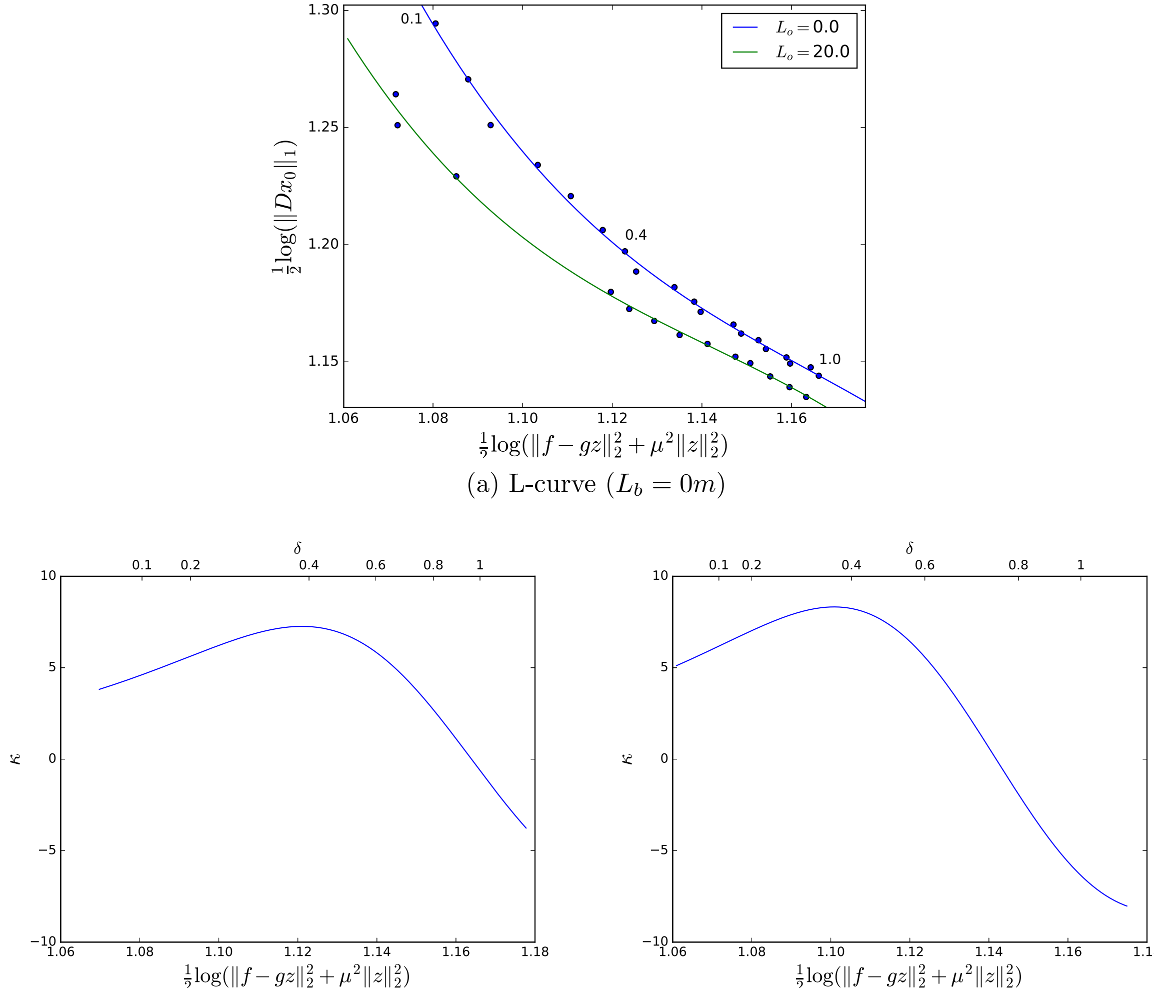
Identifying an appropriate value of the regularization parameter is an important task in a regularization problem. The method used to choose the parameter is typically problem-dependent. There are several methods that find the value of the regularization parameter assuming that the norm of the noise in observations is known [11, 49]. Conversely, there are approaches that do not need the observation noise to be known explicitly [128], [1].

Here, we follow one of the latter approaches, known as the L-curve method. In this method it is assumed that the objective function can be written in the form $\|\mathbf{y} - \mathbf{H}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_2^2$ with regularization parameter λ. The L-curve is the log-log plot of the norm of the regularized solution, $\|\mathbf{x_a}\|_2^2$, versus the norm of the corresponding residual vector, $\|\mathbf{y} - \mathbf{H}\mathbf{x_a}\|_2^2$, evaluated at the minimisation solution, $\mathbf{x_a}$. In this method, the value corresponding to the point of maximum curvature in the obtained L-shape curve is selected as the optimal regularization parameter [55, 57, 84] because at this point there is a balance between the norm of the solution vector and the fit to the observations.

For the problem studied here, a different choice of axes is made for the L-curve due to the fact that the objective function has both l_1 and l_2 terms. Recalling the minimisation problem in equation (3.7), the corresponding L-curve is defined as the plot of $\frac{1}{2}\log\|D\mathbf{x}_0\|_1$ versus $\frac{1}{2}\log\left(\|\mathbf{f} - \mathbf{G}\mathbf{z}\|_2^2 + \mu^2\|\mathbf{z}\|_2^2\right)$. In this case, the point of maximum curvature represents the balance between minimising the l_1-norm of the solution in derivative space, and the l_2-norm for the Tikkonov regularization problem.

The L-curve is shown in Figure 3.5a for $L_b = 0m$ and for two different observation error correlation length scales, $L_o = 0m$, $L_o = 20m$ where the analysis used to evaluate the functions comprising the L-curve were obtained following the data fusion experimental set-up described in Section 4.2. The corresponding curvature plots are shown in Figure 3.5b,c. Based these results, it was decided $\delta \approx 0.4$ must be chosen for the l_1-l_2 regularization method for the data fusion experiments, since this value corresponds to the maximum curvature in the L-curve.

Since in these experiments we have a known true state, the validity of the L-curve results could be verified by comparing the analysis states resulting from different δ values. Using the RMSE between the true state (taken as the data from the deformed FYI in

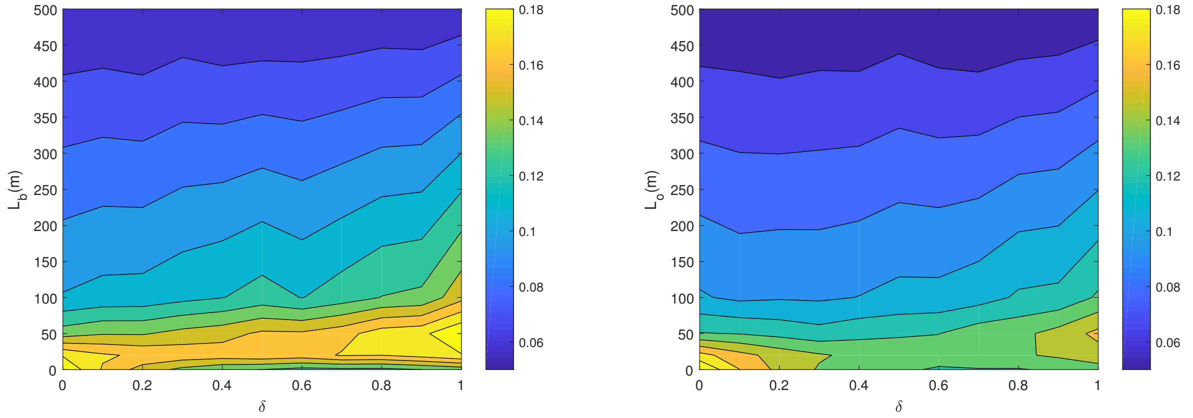


(a) L-curve ($L_b = 0m$)

(b) curvature ($L_b = 0m, L_o = 0m$)

(c) curvature ($L_b = 0m, L_o = 20m$)

Figure 3.5: Using maximum curvature of L-curve to find regularization parameter, δ. Panel (a) shows L-curve of two cases for which $L_b = 0m$. In panel (b) $L_o = 0m$ and in panel (c) $L_o = 20m$. The top axis of panels (b) and (c) shows the corresponding δ values for curvature points.



(a) $L_o = 0m$

(b) $L_b = 0m$

Figure 3.6: Analysis RMSE of the data fusion experiment over a range of δ values when (a) $L_o = 0m$ and L_b is changing and (b) $L_b = 0m$ and for a range of L_o values.

Figure 3.1) and the analyses as the error measure (see section 5.1 for further details). Two sets of experiments were carried out. In the first set, the observation error covariance matrix was diagonal ($L_o = 0m$) while L_b was varied over the range of $0m$ to $500m$, and in the second set, the background error covariance matrix was diagonal ($L_b = 0m$) and L_o was varied over the range $0m$ to $500m$. Note that the spacing of the analysis grid is $7m$, hence these length scales range from 0 to 70 times the analysis grid spacing. Results are represented in Figure 3.6. In both cases the benefit of using the l_1-l_2 approach as compared to l_2 can be seen for the shorter length scales (i.e. when L_b or L_o is lower than 50m) for δ values in the range of $0.3 < \delta < 0.7$, consistent with the results from the L-curve.

For the data assimilation experiment, similar to the data fusion experiments, δ values of $\delta = 0.01, 0.05, 0.08, 0.1, 0.2, 0.4, 0.5$ were tested. The optimal δ value was found to be 0.05, hence results using this value are shown here.

3.5 Results and Discussions

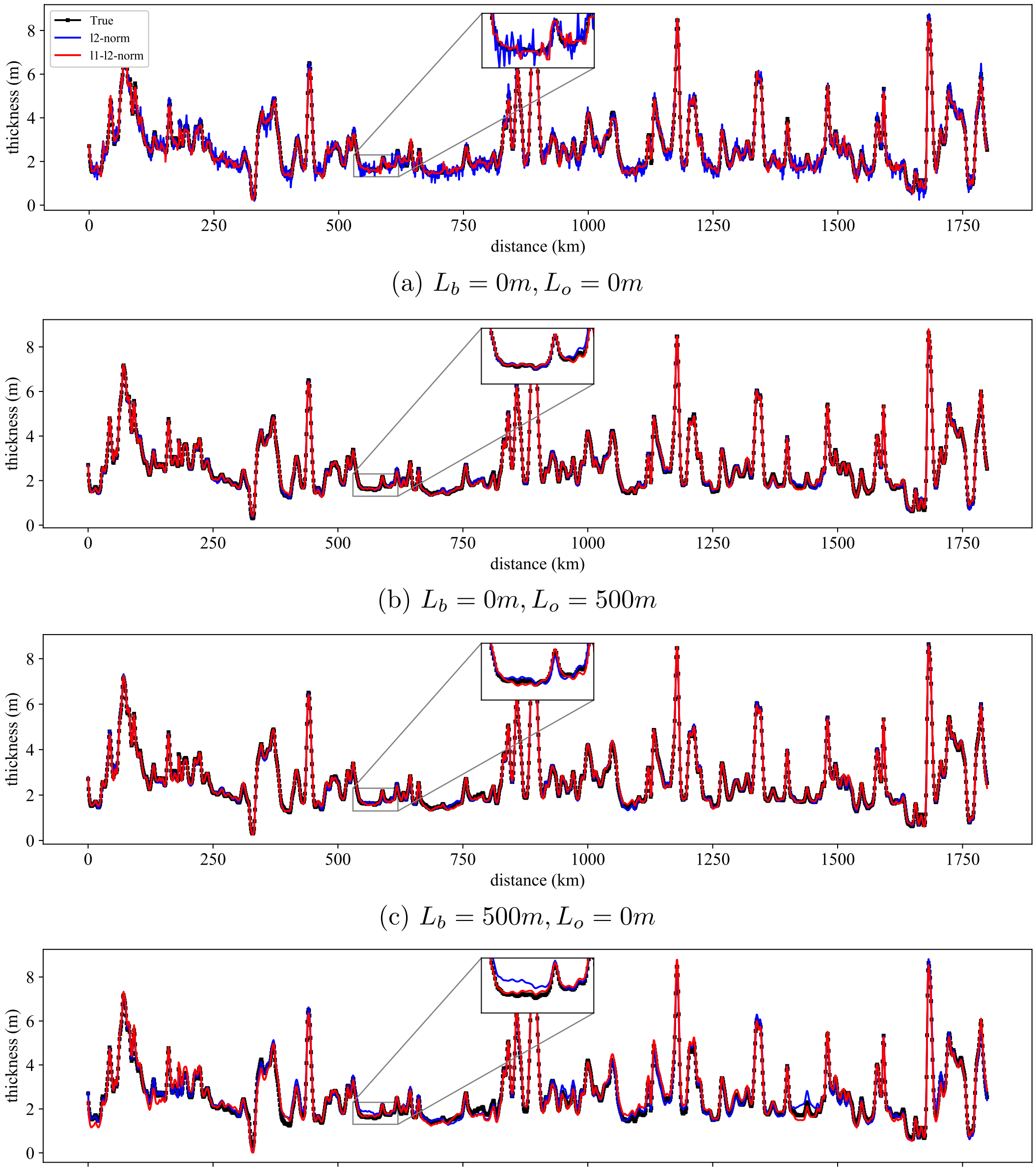
3.5.1 Results from Data Fusion Experiments

The analysis for the data fusion experiments corresponding to the case when both background and observation errors are uncorrelated is shown in Figure 3.7a. The l_1-l_2 analysis is less noisy than that from l_2 and the sharp features in the true state are captured more accurately. The analysis for the case when the background error is uncorrelated and $L_o = 500m$ is represented in Figure 3.7b. Again, in Figure 3.7b, the analysis from the l_1-l_2 method is less noisy than that from the l_2 method and the sharp features are captured more accurately. However, a benefit is no longer visibly apparent when the background error correlation length scale is $L_b = 500m$, as in Figure 3.7c and Figure 3.7d.

Table 3.2 represents the analysis errors measures described in section 5 when $\delta = 0.4$. Note the kurtosis of the true state, KURT[$\mathbf{x}_t$], and its derivative, KURT[$D\mathbf{x}_t$], were 3.12 and 10.58 respectively. Kurtosis is higher when the tails of the probability distribution are fatter which can be interpreted as sharp features in the state when kurtosis is computed on the derivative field of the state. If kurtosis of the analysis derivative is less than the true state derivative kurtosis, it shows that sharp features have been damped. Results presented in Table 3.2 confirms that when the background error is uncorrelated, the l_1-l_2-norm method outperforms the l_2-norm method in all terms of error measures. This is similar to what has been illustrated in Figure 3.7. This improvement for the MAE and RMSE measures is about 25% while the improvement is more than 40% for the ice thickness derivative fields (DIFF_ MAE and DIFF_ RMSE). In addition, the kurtosis shows that when the background error is uncorrelated, the l_1-l_2 method retains sharp features in the analysis better than l_2 method. When the observation error length scale increases, the kurtosis of both methods approaches the true state however, the l_1-l_2 method is always better. When $L_b = 20m$, the l_1-l_2 method is slightly better than the l_2 method regardless of the observation error correlation length scale. However, when the background error correlation length scale is $L_b = 500m$, the resulting analysis states are relatively smooth and DIFF_ MAE and DIFF_ RMSE are almost zero for all cases. Finally, it should be noted that while the kurtosis of the analysis states for both methods are similar, in all cases kurtosis of the analysis derivative is always better for the l_1-l_2-norm method, which shows the ability of this method in retaining the sharp features in ice thickness.

3.5.2 Results from Data Assimilation Experiments

The root-mean-squared (RMS) ice thickness errors from the data assimilation experiments are shown in Figure 3.8 for various combinations of background and observation error correlation length scales. In all cases the errors are reduced when l_1-l_2 is used as compared with l_2. Time traces of the data assimilation states (see Figure 3.13 and Figure 3.14) showed that this growth in errors, which is also seen in the ice concentration and ice velocity (Figure 3.9) is related to the development of openings in the ice cover, which are better captured with l_1-l_2 than l_2.



(a) $L_b = 0m, L_o = 0m$

(b) $L_b = 0m, L_o = 500m$

(c) $L_b = 500m, L_o = 0m$

(d) $L_b = 500m, L_o = 500m$

Figure 3.7: Data fusion analysis states for l_2-norm (blue), l_1-l_2-norm (red) and true state (black) for different background and observation error correlation length scales for $\delta = 0.4$.

Table 3.2: Data fusion analysis errors for different background and observation error correlation length scales (L_b and L_o) when $\delta = 0.4$, KURT[$\mathbf{x}_t$] $= 3.12$ and KURT[$D\mathbf{x}_t$] $= 10.58$. The best results are in bold font.

$L_b(m)$	$L_o(m)$	regularization method	MAE	RMSE	KURT[$\mathbf{x}_a$]	DIFF_ MAE	DIFF_ RMSE	KURT[$D\mathbf{x}_a$]
0	0	l_2-norm	0.16	0.20	3.00	0.22	0.28	1.69
		l_1-l_2-norm	**0.10**	**0.14**	**3.06**	**0.11**	**0.15**	**8.46**
	20	l_2-norm	0.14	0.18	3.02	0.15	0.19	3.8
		l_1-l_2-norm	**0.11**	**0.14**	**3.08**	**0.09**	**0.12**	**8.67**
	500	l_2-norm	0.05	0.06	3.11	0.01	0.01	10.54
		l_1-l_2-norm	**0.04**	**0.05**	3.11	0.01	0.01	**10.60**
20	0	l_2-norm	0.15	0.19	3.01	0.17	0.21	2.95
		l_1-l_2-norm	**0.13**	**0.17**	**3.04**	**0.15**	**0.19**	**4.08**
	20	l_2-norm	0.16	0.20	2.94	0.15	0.18	3.71
		l_1-l_2-norm	**0.14**	**0.18**	**2.95**	**0.13**	**0.16**	**5.11**
	500	l_2-norm	0.06	0.07	**3.13**	0.00	0.01	10.60
		l_1-l_2-norm	0.06	0.07	3.15	0.00	0.01	**10.58**
500	0	l_2-norm	0.05	0.06	3.14	0.00	0.00	**10.59**
		l_1-l_2-norm	0.05	0.06	**3.10**	0.00	0.00	10.51
	20	l_2-norm	0.06	0.08	3.08	0.00	0.00	10.55
		l_1-l_2-norm	0.06	0.08	3.08	0.01	0.00	**10.60**
	500	l_2-norm	0.15	0.19	**2.96**	0.01	0.00	10.50
		l_1-l_2-norm	0.15	0.19	2.93	0.01	0.00	**10.57**

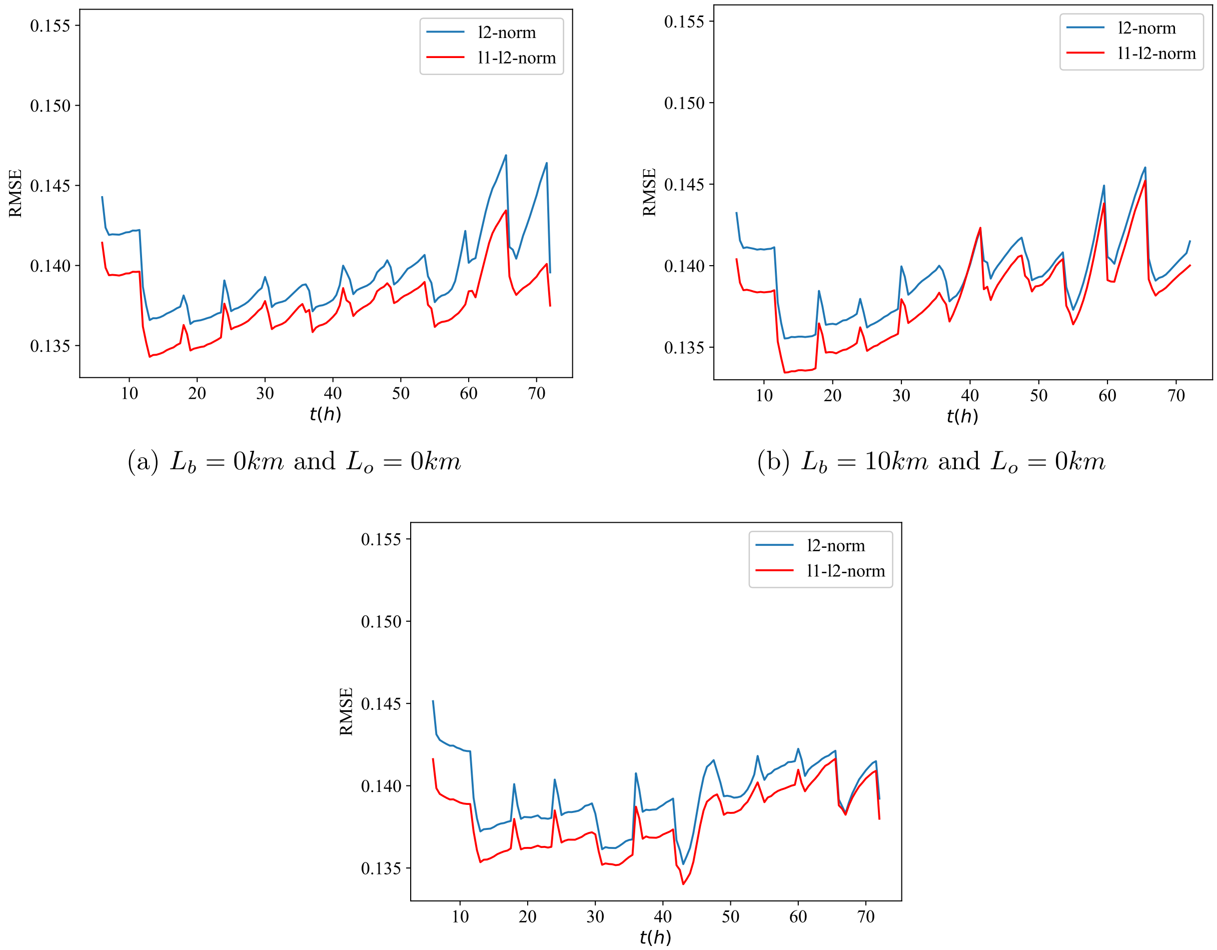
Histograms of the analysis states (ice thickness) and their derivative fields at each analysis time are presented in Figure 3.10 and Figure 3.11 respectively for the case where $L_b = 0km$ and $L_o = 0km$. In both figures, the first row illustrates the histogram of the true state. In addition, for better comparison, the histogram of the true state is overlayed on the histograms of the data assimilation method. In Figure 3.10 (top row) it is difficult to see substantial differences between the ice thickness histograms, although the derivative fields in Figure 3.11, show that for the l_1-l_2 method, there is a higher level of sparsity on the derivative field of analysis as compared to the l_2 method. This is consistent with the expected results, since the presence of the l_1 term on the ice thickness derivative should result in more zeros in this field.

A representative result from the data assimilation experiments is shown in Figure 3.12. This result is obtained at $t = 64h$ for L_b=0km and L_o=0km. Figure 3.12 shows that the reduction in ice concentration at spatial index 389 is captured more accurately for l_1-l_2 as compared to l_2. The time trace of model states for this spatial index is shown in Figure 3.13. Note the states are in agreement when the ice is thick and the concentration is high, but differences develop as the ice thins and the concentration drops. A similar result is shown in Figure 3.14, where the reduction in ice concentration and thickness, is better captured with the l_1-l_2 regularization. Note also that both time traces show the ice velocity is also in better agreement with the true state for l_1-l_2.

Similar results were found for other background and observation error length scales and also for the case when the ice cover was consolidated (although these events were observed less frequently for the consolidated ice cover). In general, openings in the ice cover and related changes in the ice velocity are better captured with l_1-l_2 than l_2. This may be due to the relationship between ice thickness, ice strength and bulk viscosity, and the role this plays in the momentum equation (see Appendix B). The exact details of this interaction are complicated and will depend on the chosen parameters and numerical methods used. This will be investigated in a future study.

3.6 Summary

In this chapter it was demonstrated that sea ice thickness exhibits a sparse representation in the derivative domain. This has been demonstrated using (i) sea ice thickness measurements from an airborne electromagnetic (AEM) sensor over the Beaufort Sea; (ii) submarine upward looking sonar data; and (iii) sea ice thickness from Cryosat. To retain this feature when using sea ice thickness data in a data fusion or data assimilation scheme, the use of an additional term in the objective function to constrain sparsity on

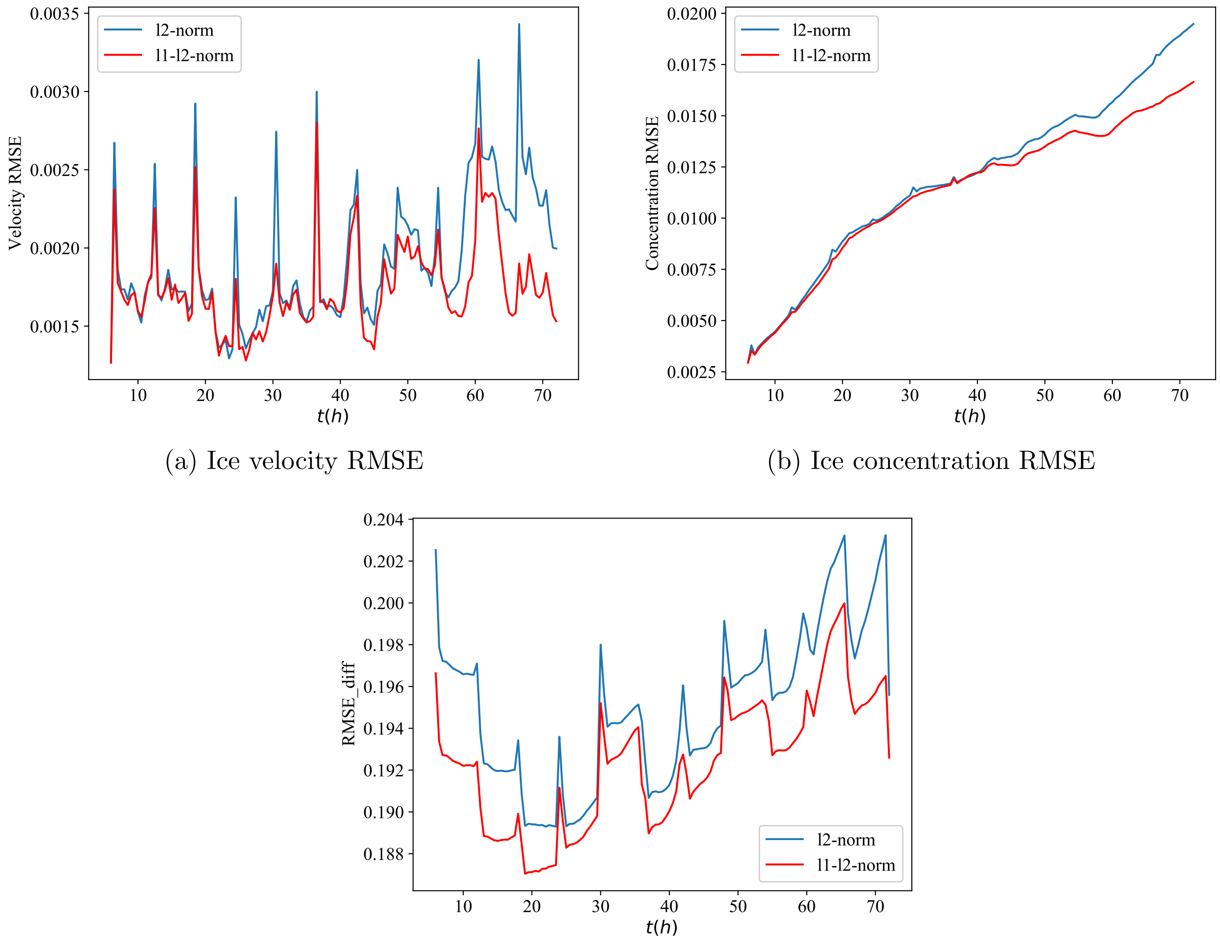


(a) $L_b = 0km$ and $L_o = 0km$

(b) $L_b = 10km$ and $L_o = 0km$

(c) $L_b = 0km$ and $L_o = 10km$

Figure 3.8: RMSE of the data assimilation ice thickness [m] states when (a) $L_b = 0km$ and $L_o = 0km$, (b) $L_b = 10km$ and $L_o = 0km$ and (c) $L_b = 0km$ and $L_o = 10km$ for the 72 hour data assimilation experiment. Data are assimilated every 6 hours.



(a) Ice velocity RMSE

(b) Ice concentration RMSE

(c) Derivative of ice thickness RMSE

Figure 3.9: RMSE of the (a) ice velocity, (b) ice concentration and (c) ice thickness derivative when $L_b = 0km$ and $L_o = 0km$ for the 72 hour data assimilation experiment. Data are assimilatied every 6 hours.

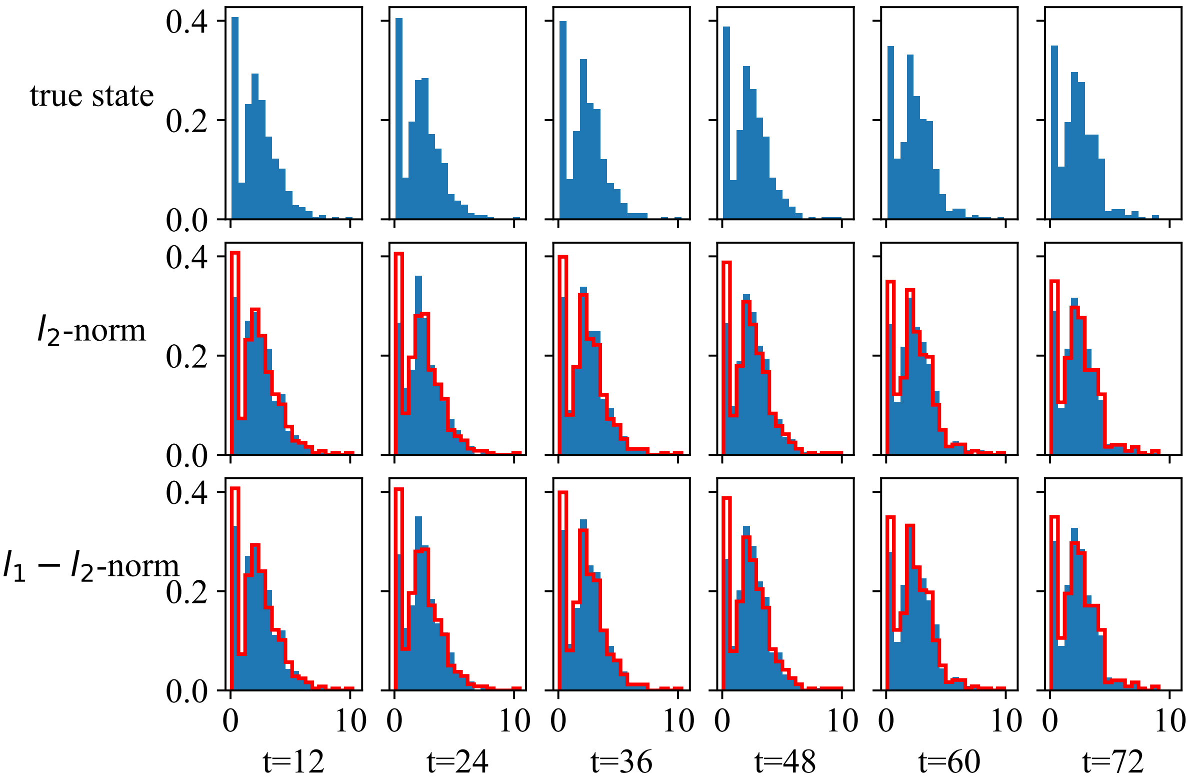


Figure 3.10: Histograms of the analysis states (ice thickness [m]) from the data assimilation when $L_b = 0km$ and $L_o = 0km$. Each column shows the analysis state at the indicated time. The top row shows the true model, and the second and third rows show l_2-norm and l_1-l_2-norm results respectively. Histogram from the true state is overlayed by red colour in the second and third rows.

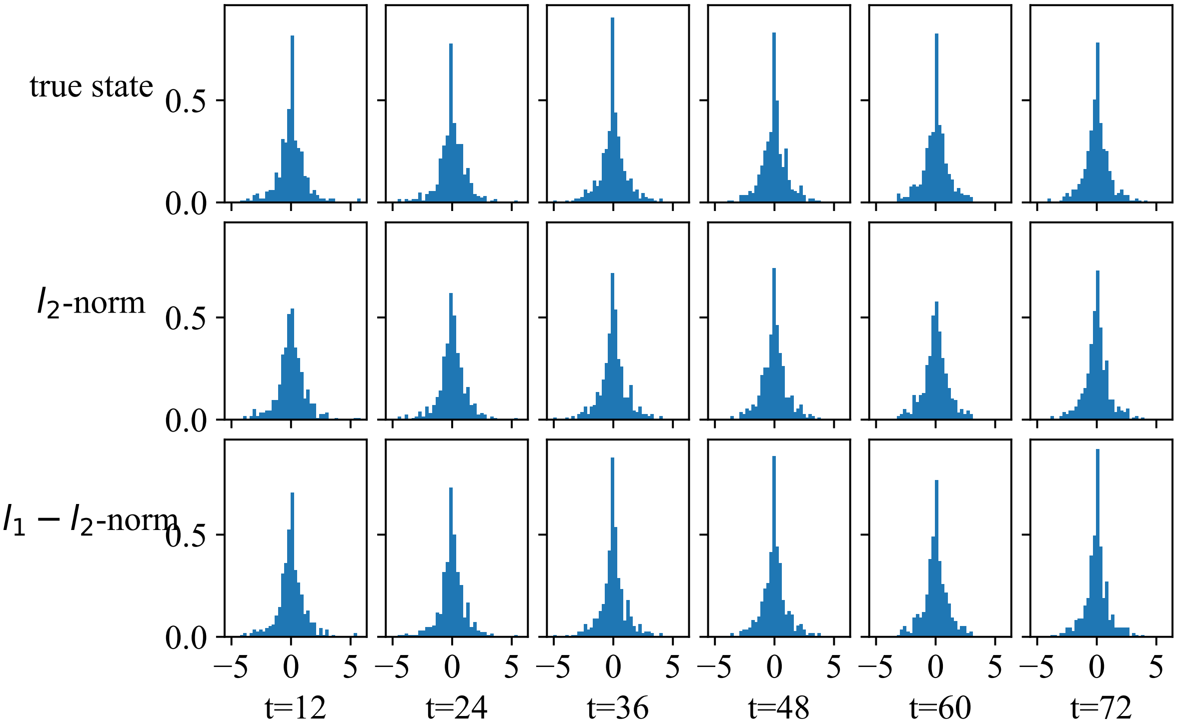


Figure 3.11: Histograms of the derivative of analysis states (ice thickness derivatives [-]) from data assimilation when $L_b = 0km$ and $L_o = 0km$. Each column shows a different time step of the model. The top row shows the true model, and the second and third rows are showing l_2-norm and l_1-l_2-norm results respectively.

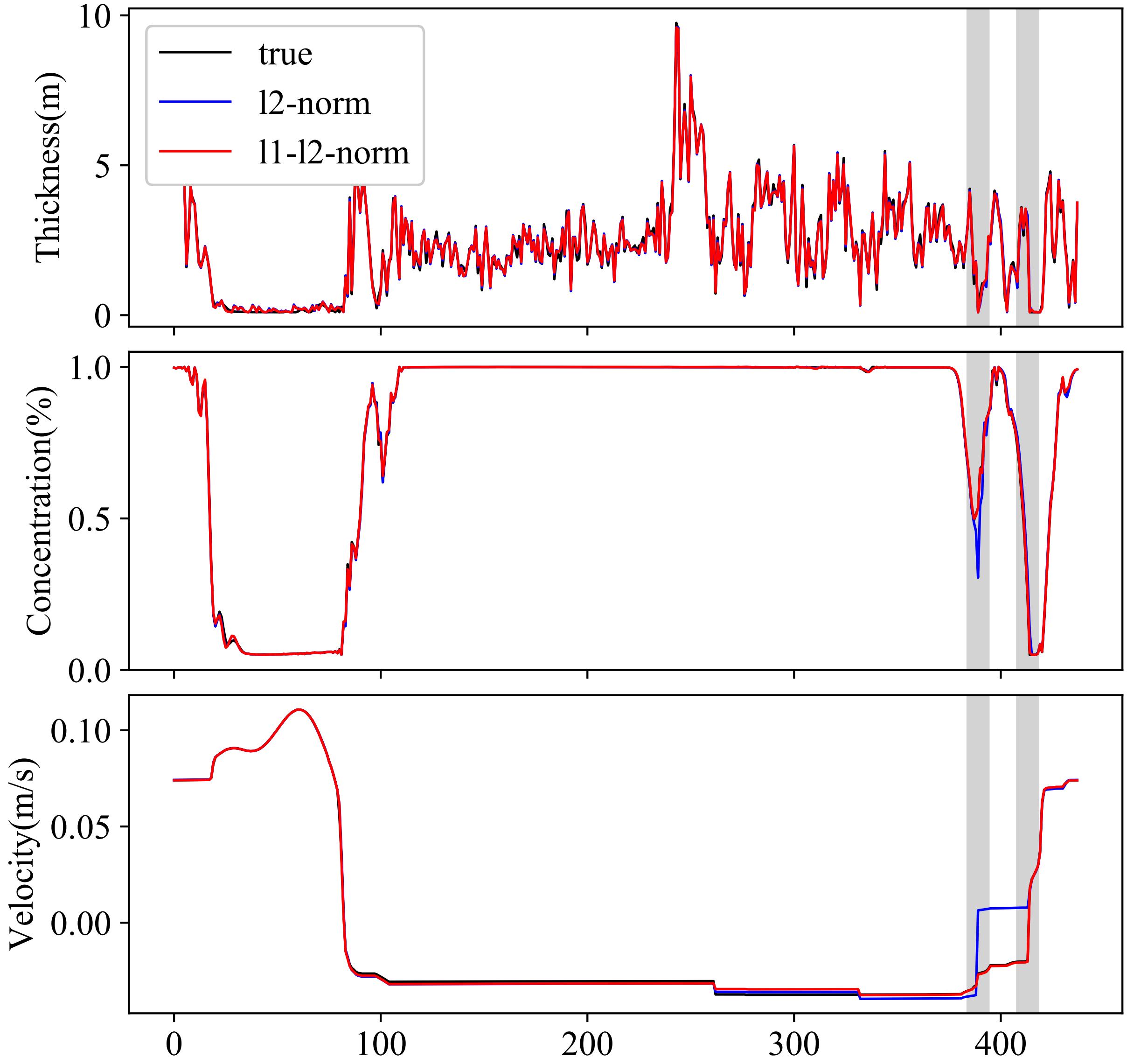


Figure 3.12: An example of data assimilation states at t=64.0(h) when $L_b = 0km$ and $L_o = 0km$. The shaded regions indicate the locations at which time traces (shown in Figure 3.13 and Figure 3.14) are taken.

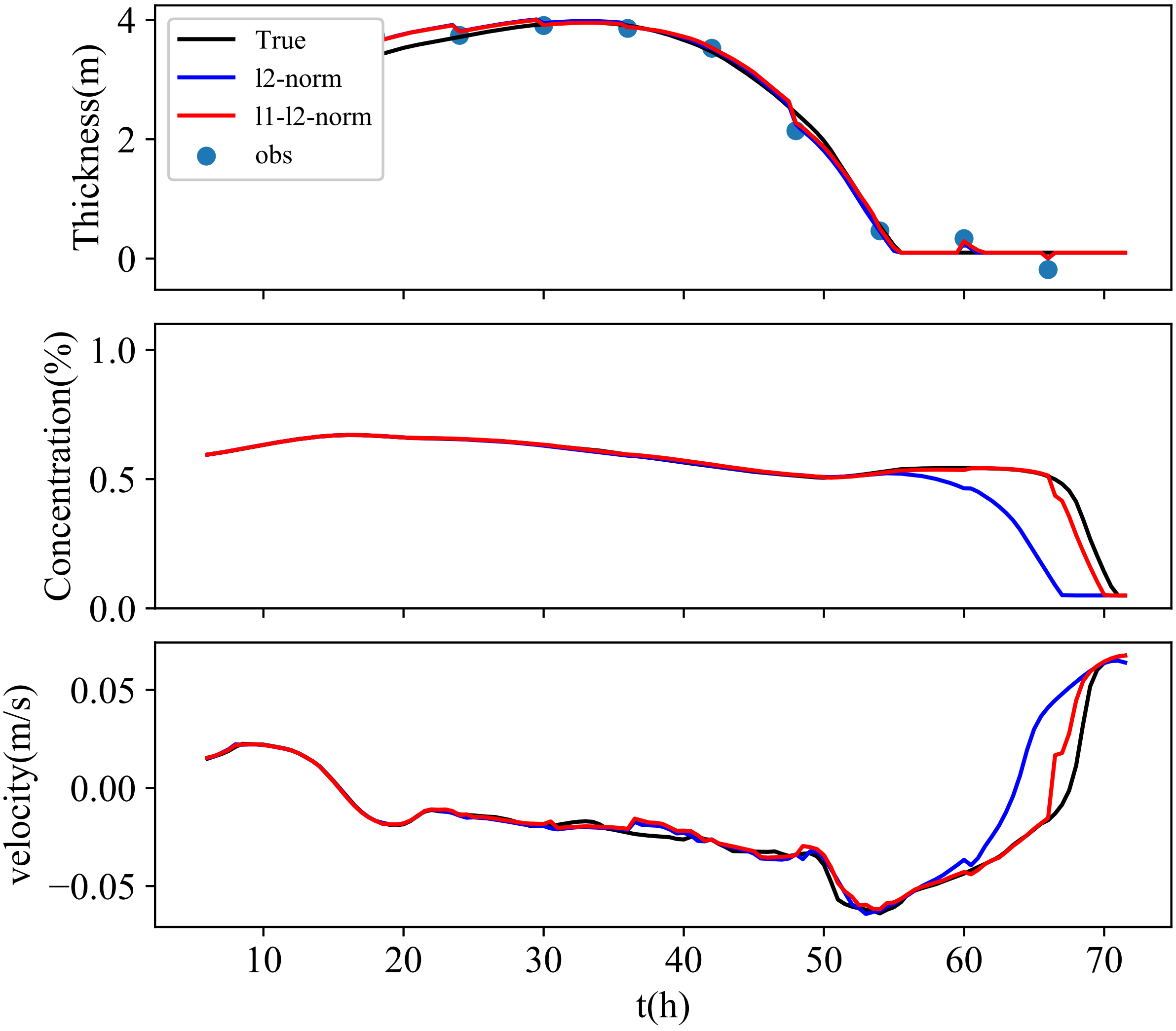


Figure 3.13: Model states from a single realisation of the data assimilation experiment for $L_b = 0km$ and $L_o = 0km$ at index=389 (indicated by vertical line in Figure 3.12) The decrease in ice concentration and thickness is overestimated for l_2.

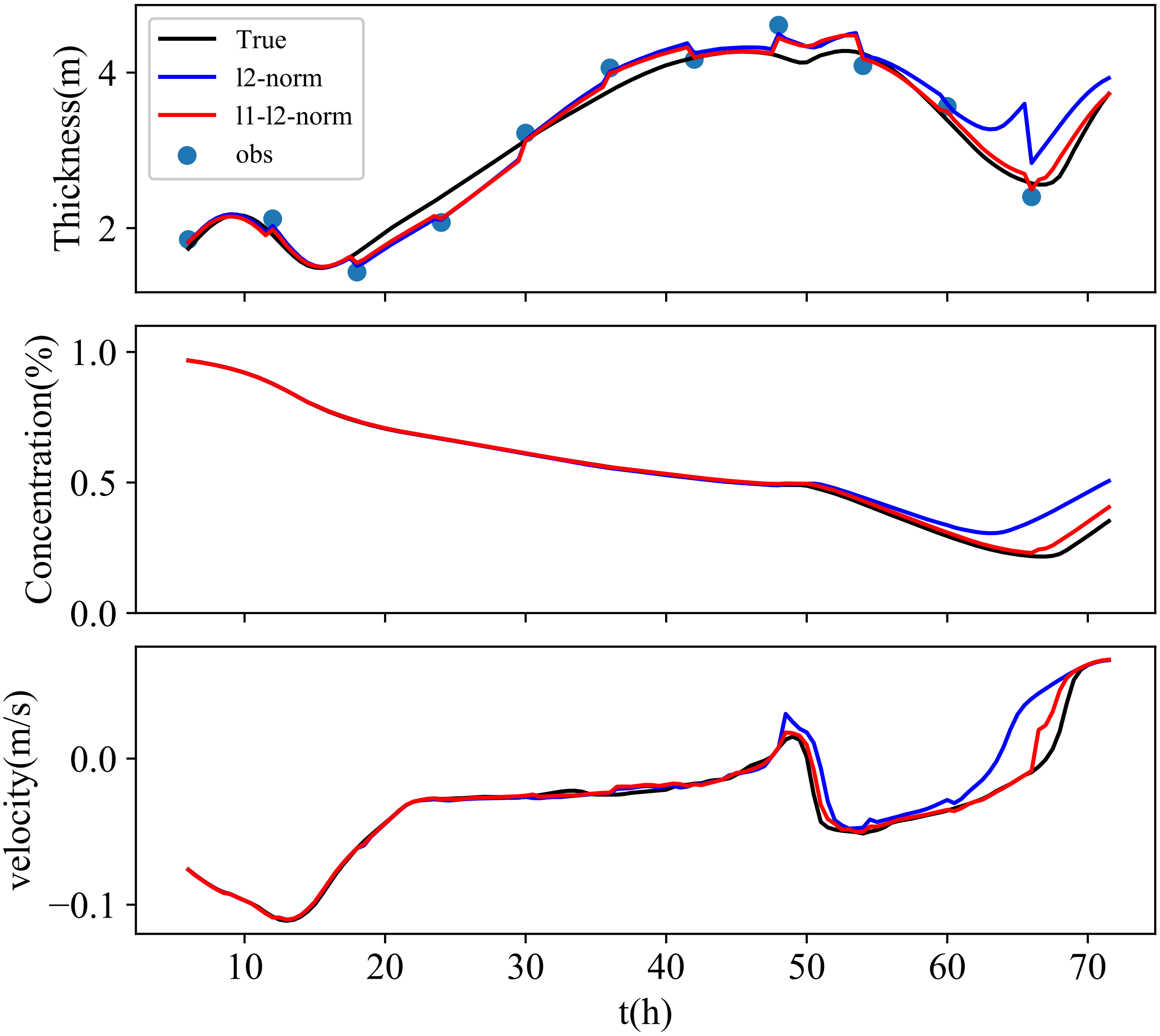


Figure 3.14: Model states from a single realisation of the data assimilation experiment for $L_b = 0km$ and $L_o = 0km$ at index=245 (indicated by vertical line in Figure 3.12). The opening in the ice cover is predicted better for the l_1-l_2 method than for the l_2 method.

the derivative of the ice thickness state, is evaluated. This l_1-l_2 formulation is compared with the standard l_2 regularization first using data fusion, and then by carrying out data assimilation experiments using a toy sea-ice model.

For data fusion a clear benefit to the l_1-l_2 formulation is seen when the background correlation error length scale is small (on the order of twice the analysis grid spacing). Based on previous studies [17], it can be expected that in the vicinity of a sharp feature (e.g. ice edge) the background error correlation length scale may be in this range. This data fusion result could be relevant for the generation of merged sea ice products, where sharp features are desired (see for example the ice concentration used at the lower boundary in [7]).

For data assimilation a clear benefit is also seen to the l_1-l_2 regularization, although the impact of the error correlation length scales on the difference between the l_1-l_2 method and the l_2 method is less clear. This may be due to the spatial averaging of ice thickness that was required to increase the scale of the data, or it could be due to the model dynamics. However, based on the results shown here, it can be noted that the l_1-l_2 method is superior with regards to capturing openings in the ice cover than the conventional l_2 method. This was observed for a variety of error correlation length scales, values of the regularization parameter (μ), and model initial conditions.

Chapter 4

Ice/Water Classification Using Neural Networks

4.1 Introduction

In this chapter, a neural network approach is proposed for the task of ice/water classification from a dataset consisting of six years of SAR images. The chapter starts by reviewing the concept of neural networks followed by the details of the proposed method. The results of training and testing the method on pure ice and water samples as well as samples with different ice concentrations are discussed. In addition, the performance of the neural network approach was investigated for the case when the training set is limited to the samples of one year. The results are compared with the logistic regression probability model that has been recently proposed for the same dataset [76].

4.2 Neural Network Approach

Artificial neural networks (ANN) are a powerful machine learning tool. They have been widely used to solve remote sensing problems, including sea ice applications, over the past 20 years. Neural networks are classified in the category of nonparametric machine learning methods i.e., they do not require any specific assumption about the statistical distribution of the data used for training.

An artificial neural network is constructed by elements called artificial neurons. Each neuron receives an input and produces an output using their specified activation function.

The network is constructed by feeding the outputs of one layer of neurons, as input, to the next layer of neurons. Each connection in this network has a weight. For solving the ice/water classification problem, multilayer perceptron (MLP), a form of feedforward neural networks which is one of the most widely used ANN models [112], is used in this thesis. An MLP is typically trained using the error backpropagation algorithm [59].

An MLP consists of three types of neuron layers: input layer, hidden layers and output layer. Figure 4.1a illustrates an MLP with 2 hidden layers. Having a training sample $(\mathbf{x}, \mathbf{y})$ where input $\mathbf{x}$ has n features, $\mathbf{x} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$, and $\mathbf{y}$ is the sample's label, the input layer is constructed by associating one neuron to each feature. Each given input vector propagates through the network layer by layer until it reaches the output layer. If the value of neurons at layer l with m neurons are represented by $a_i^l, i = 1, 2, ..., m$, the value of neuron j of the next layer, a_j^{l+1}, is calculated as,

$$a_j^{l+1} = \psi \left(\sum_i^m a_i^l W_{ij}^l + b^l \right). \tag{4.1}$$

This process is also shown in Figure 4.1b. In (4.1), W_{ij}^l is the weight of connection between neuron i and j, ψ is the activation function, and b^l is the bias of layer l added to the weighted sum. When the propagation gets to the output layer, the output of the network, $\hat{\mathbf{y}}$, is compared to the desired target, $\mathbf{y}$, using a loss function and the error is calculated. The resulting error is propagated back through the network until error associated with each neuron in the hidden layers of the network has been calculated. Based on these calculated errors, the value of each weight in the network is updated. This optimization technique repeatedly performs the propagation process followed by weight update, until a stopping criterion is satisfied.

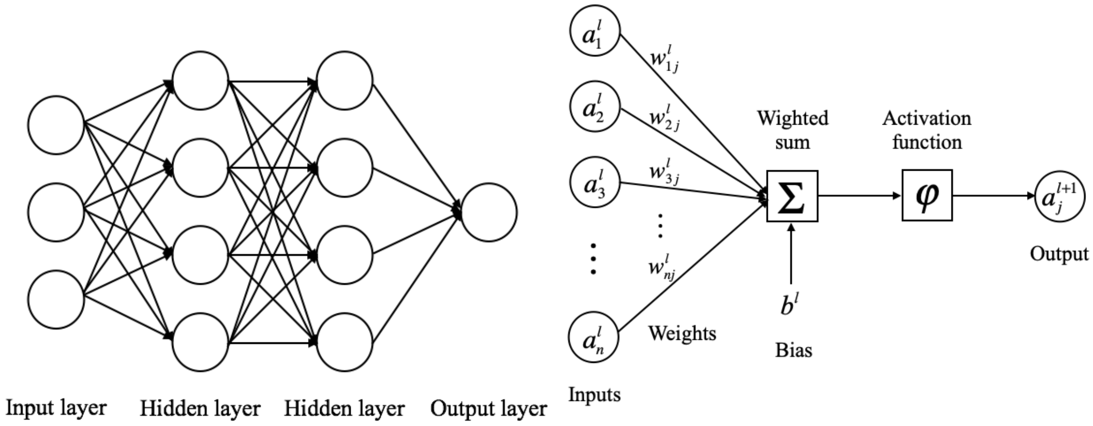
Except for input layer, each neuron of an MLP is associated with an activation function that maps the result of weighted sum to an output value for the next layer. This activation function can be used to limit the amplitude of the output of a neuron. Sigmoid and rectified linear unit (ReLU) are the two most commonly used activation functions. The sigmoid function is defined as

$$f(x) = \frac{1}{1 + e^{-x}}, \tag{4.2}$$

and the ReLU function is

$$f(x) = \max(0, x). \tag{4.3}$$

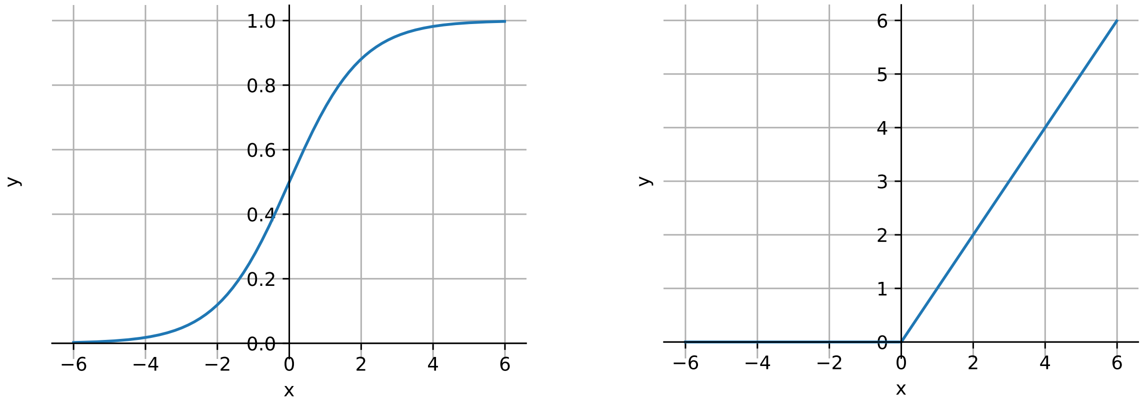
The sigmoid function has domain of all real numbers, and generates outputs from 0 to 1 (Figure 4.2a) which is suitable for predicting probability as an output. The ReLU function



(a) MLP

(b) forward propagation

Figure 4.1: A typical multilayer perceptron (MLP). (a) Architecture of an MLP with 2 hidden layers. (b) An example of forward propagation in an MLP showing how neurons of layer l generate neuron j of the next layer.



(a) Sigmoid function

(b) ReLU function

Figure 4.2: Two samples of activation functions. (a) Sigmoid function, (b) ReLU function.

is known as a sparse activation (Figure 4.2b) by turning all negative values to zero so they do not propagate through to the next layer.

To calculate the error of the NN output, regression problems usually use a mean squared error (MSE) loss function defined as

$$L(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{y}_i - \hat{\mathbf{y}}_i)^2, \tag{4.4}$$

where $\hat{\mathbf{y}}_i$ is the output of neuron i in the output layer, y_i is its corresponding desired target and n is the number of target labels equal to the number of neurons in the output layer. In contrast, for the binary classification problems, binary cross-entropy is commonly used, which is defined as

$$L(\mathbf{y}, \hat{\mathbf{y}}) = -\frac{1}{n} \sum_{i=1}^{n} [\mathbf{y}_i \log \hat{\mathbf{y}}_i + (1 - \mathbf{y}_i) \log (1 - \hat{\mathbf{y}}_i)]. \tag{4.5}$$

One popular optimization approach for updating network weights is gradient descent. Using this approach, each weight is updated using the partial derivative of the loss function with respect to the weight as

$$W_{ij}^{l} = W_{ij}^{l} - \eta \frac{\partial L(\mathbf{y}, \hat{\mathbf{y}})}{\partial W_{ij}^{l}}, \tag{4.6}$$

where W_{ij}^{l} is the weight of connection between neuron i in layer l to neuron j in layer $l+1$, η is the learning rate and $\partial L(\mathbf{y}, \hat{\mathbf{y}})/\partial W_{ij}$ is the loss function gradient with respect to weight point W_{ij}^{l}. Equation (4.6) shows that the learning rate is a parameter that controls the weights adjustments with respect the loss gradient. Using the chain rule the partial derivative of equation (4.6) can be simplified as

$$\frac{\partial L(\mathbf{y}, \hat{\mathbf{y}})}{\partial W_{ij}^{l}} = \delta_j^{l+1} a_i^{l} \tag{4.7}$$

with

$$\delta_j^{l+1} = \begin{cases} (a_j^{l+1} - \mathbf{y}_j) f'(\mathbf{x}_j) & \text{if } l+1 \text{ is the output layer} \\ \left(\sum_{k=1}^{n} \delta_k^{l} W_{kj}^{l}\right) f'(\mathbf{x}_j) & \text{if } l+1 \text{ is not the output layer} \end{cases} \tag{4.8}$$

where k is the number of neurons in layer l and recalling that a_j^{l} is the value of neuron j at layer l, and $f(\mathbf{x}_j)$ and $f'(\mathbf{x}_j)$ are the outputs of activation function and its derivative on neuron j. It should also be noted that, in (4.6), $\eta \in [0,1]$ is a small positive constant that controls the size of step.

Finally, there are three different modes of training in the neural network training process: batch mode, mini-batch mode, and stochastic mode. In all modes, one complete pass through the whole dataset is called one epoch. In the batch mode, network weights are updated after the whole dataset is passed through the network. In this case, the loss function is calculated based on the accumulated values from the whole dataset. In mini-batch mode, a mini-batch of stochastically selected samples of the dataset is selected and network weights are updated after each mini-batch has completely passed through the network. In the stochastic learning, network weights are updated after each single training sample is represented to the network. Among these modes, the batch training has the higher chance to find the global minimum but it is very slow and loading the whole dataset is not always feasible, especially for large datasets. Stochastic learning is the fastest one but the loss updates do not always decrease from iteration to iteration and have high variance. In addition, the final accuracy for this method is usually the lowest. Finally, the mini-batch mode has reduced loss update variance and allows parallel computations, which makes it very fast. However, batch size (number of samples per batch) is a hyper-parameter that has to be defined.

For the ice/water classification problem, using MLP with different numbers of hidden layers is investigated here. The sigmoid function is selected as the activation function of the output layer while for hidden layers ReLU is selected. The hyperbolic tangent is also used as an activation function of the output layer in some cases which is a shifted/scaled version of the sigmoid function. For the optimization approach, Adam optimizer [75] is chosen. Adam optimizer is similar to gradient descent but it considers an individual adaptive learning rate for each parameter. The algorithm estimates first-order moment and second-order moment of the gradient using exponential moving average, and corrects the bias in these estimates. The final weight update is proportional to learning rate times first-order moment divided by the square root of second-order moment. This characteristic makes Adam a computationally efficient optimizer for problems with a large dataset or a large model. Adam takes three hyperparameters: the learning rate, the decay rate of first-order moment, and the decay rate of second-order moment [111].

4.3 Data

This section is dedicated to describe the dataset developed by Komarov et. al. [76], which is employed in this study.

Data Sources

The training and test features of the utilized dataset were extracted from RADARSAT-2 ScanSAR HH-HV images and global environment multiscale model (GEM) regional deterministic forecasts of Environment and Climate Change Canada (ECCC). The verification data were collected from corresponding CIS image analysis products. A detailed description of each source is provided in [76]. Details related to the present study are briefly presented.

The RADARSAT-2 collection contains 15405 dual-pol images covering Alaskan, Canadian, and West Greenland waters from November 1, 2010 to September 30, 2016. Each image has a nominal spatial resolution of 50m×50m and covers a spatial region of approximately 500km×500km with incidence angles from 20° to 50° [97]. Figure 4.3 shows the covered regions with colors representing the month of SAR image acquisition.

The CIS image analyses are developed manually by ice specialists at the CIS. Since these products are generated mainly based on RADARSAT-2 images, there is no time difference between the images used in the dataset and the image analysis. Other sources of information that might be used in the development process of image analysis based on their availability include optical satellite images and visual observations acquired by ships and aircrafts. Each image analysis is generally composed of contiguous polygons where each polygon contains information about the total ice concentration, partial ice concentration, stage of development and form of ice [97]. The reported ice concentrations (total and partial) are quantized with 10% increment. Here, the dataset used 5km resolution raster version of image analysis products. These manual ice analyses are associated with observational and mapping errors [23]. Observational errors are mainly caused by ice specialists mistakes due to lack of time while mapping errors are relevant to drawing and generalization errors in the production. However, CIS image analysis are still the most popular data source for quality assessment in ice concentration studies (e.g., [115, 130]).

Another source of information in the dataset is the 10m wind speed produced by the regional version of the ECCC operational numerical weather prediction (NWP) using the GEM model. The model wind speeds are produced every hour at ∼10km spatial resolution.

Dataset Features

To collect samples from SAR data, HH and HV images of RADARSAT-2 products were first radiometrically calibrated and then a median filter of size 3×3 was applied to reduce the speckle noise. Next, windows of size 41×41 pixels or $2.05km \times 2.05km$ were selected for

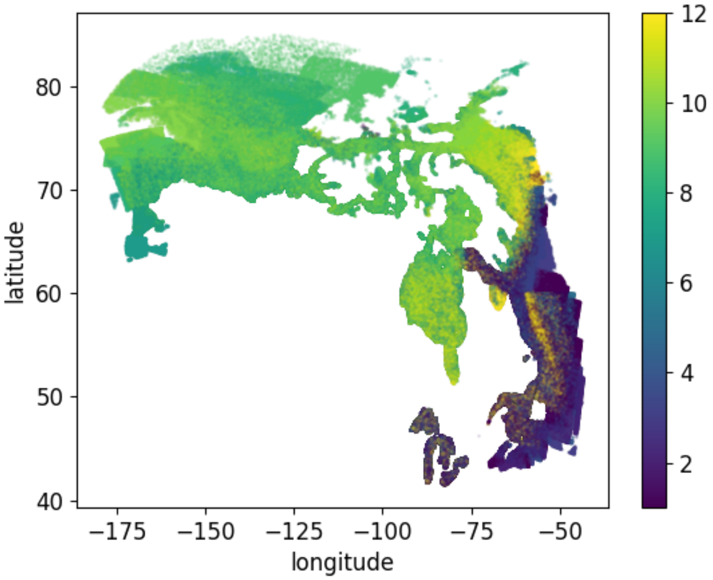


Figure 4.3: Map of region over which the SAR database samples were acquired. Colors indicating the month of acquisition of the SAR imagery with 1-12 representing the months sequentially from January to December. Due to the large number of images, samples were thinned for clarity. The distribution of samples indicates that in the winter months most images were acquired on the east coast of Canada, whereas in the summer they acquired at higher latitudes.

feature extraction. Samples with land or image boundary within 161×161 pixels ($8.05km \times 8.05km$) are not included in the dataset.

In this database, the following four features are extracted from the 41 × 41 sampling windows:

1. SAR wind speed: For each pixel of sample window in SAR images, first the wind speed is calculated using an approach proposed by Komarov et. al. [79] and then, the average wind speed over the window is computed. The SAR wind speed calculated based on this approach, is only valid over open water areas.

2. NWP wind speed: For each sampling window the corresponding NWP wind speed is interpolated based on the location. Since the SAR wind speed is invalid over ice regions, the difference between the SAR and NWP wind speed will be close to zero

Table 4.1: Number of samples in each subset of ice/water dataset. From the time period between November 1, 2010 and September 30, 2016, the samples of year 2013 where used for testing and the rest for training.

Ice concentration [%]	Samples in training set	Samples in test set
0	6,325,869	1,490,240
100	4,871,680	990,638
>0 & <100	27,510,377	5,984,659

around open water areas and different from zero in ice-covered areas.

3. HH-HV spatial correlation: This correlation is calculated as:

$$c_{HH-HV} = \frac{\mathrm{cov}(\sigma^0_{HH}, \sigma^0_{HV})}{[\mathrm{cov}(\sigma^0_{HH}, \sigma^0_{HH})\mathrm{cov}(\sigma^0_{HV}, \sigma^0_{HV})]^{0.5}} \tag{4.9}$$

where $\mathrm{cov}(\sigma^0_{HH}, \sigma^0_{HV})$ is the covariance between the HH and HV backscatters of the sampling window.

4. Standard deviation of SAR wind speed: This is another feature calculated over each sampling window. The spatial distribution of wind speeds should not exhibit significant variability over the 2.05km×2.05km scale, hence this feature is supposed to be relatively small over open water. Even though the calculated SAR wind speed is not valid over ice-covered areas, it is expected that this feature has large variability over those regions which should help to distinguish ice from water.

To allow direct comparison with the previous study on this dataset [76], the same training and test sets are used here. The samples of year 2013 were selected for the test set and the remaining samples were used for training. Availability of samples from six years, different seasons and regions, and various weather conditions in the training set, make the dataset very comprehensive. Table 4.1 shows the number of samples for each of training and test set for each ice concentration category.

4.4 Results and Discussion

The performance of the neural network approach discussed in Section 4.2 is evaluated on the dataset described in Section 4.3. For that purpose, the neural network setup is described first.

4.4.1 Neural Network Setup

To investigate the impact of MLP architecture on the ice/water classification, three different network architectures are examined: shallow network (1 hidden layer with 100 neurons), mid-size network (5 hidden layers with 100-500 neurons in each layer), and a deeper network (10 hidden layers with 100-500 neurons in each layer). The number of neurons in each layer were set empirically, however, there are several studies on how to set the number of neurons in hidden layers [117]. In addition, all architectures are evaluated with three and four input features. The logistic regression method proposed by Komarov used three features including: SAR-NWP wind speed, HH-HV spatial correlation, and standard deviation of SAR wind speed. The neural network approach trained on three features uses the same features. However, the four feature network uses SAR wind speed and NWP wind speed as two separate features instead of using their difference. Table 4.2 represents the values of MLP parameters. It should be noted that during the initial experiment, batch size of 2048 and 256 were both tested but based on time efficiency and classification accuracy, batch size of 2048 was fixed for all the models. The training data set was split further into two sets, with 70% of the data used for training the NN directly (i.e., used in minimization of the loss function) and 30% used as validation data. The training procedure was stopped if the validation loss does not decrease after five epochs or the maximum number of epochs, which is 30, is reached.

For measuring the accuracy of each model, six statistical parameters are evaluated and reported: 1) training loss; 2) rate of correctly classified pure ice samples denoted as ice accuracy, 3) water accuracy, as the rate of correctly classified water samples; 4) total accuracy, derived by rate of correctly classified samples from both classes; 5) total misclassified; and 6) rate of unknown samples.

4.4.2 Results on Pure Ice and Water Samples

As the first set of experiments, the three different MLP models were trained once with three features and once with four features. Figure 4.4 shows the training and validation set loss value over the 30 epochs for the MLP with four features and one hidden layer. The loss curves are similar for other network configurations.

After training, the model accuracy was tested against the test dataset which consists of pure ice and water samples of year 2013. Labeling of the samples based on the predicted

Table 4.2: Parameters of the MLP approach with their values.

MLP parameters	Parameter value
Number of input features	3,4
Number of hidden layers	1,5,10
Loss function	Binary cross-entropy
Batch size	2048
Optimizer	Adam optimizer
Initial learning rate	0.001
Decay rate of 1st-order moment	0.9
Decay rate of 2nd-order moment	0.999
Number of epochs	30

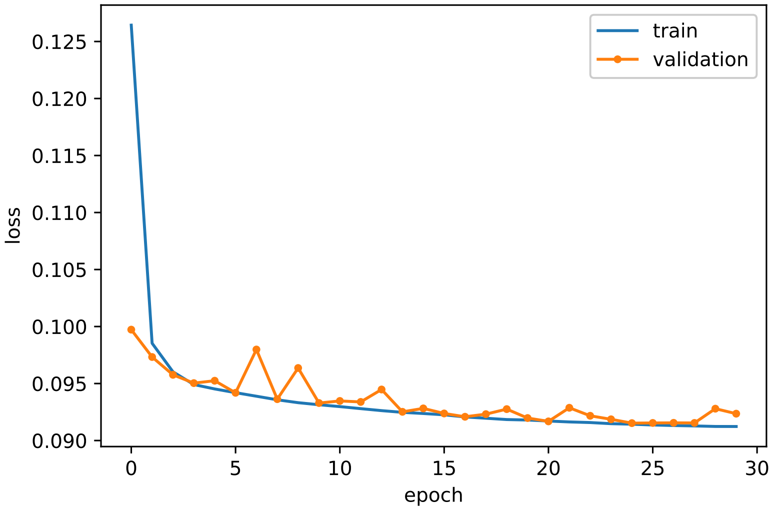


Figure 4.4: The value of cross-entropy loss function for training and validation set at each epoch during the training process of an MLP with four features and one hidden layer. The MLP training is stopped after 30 epochs.

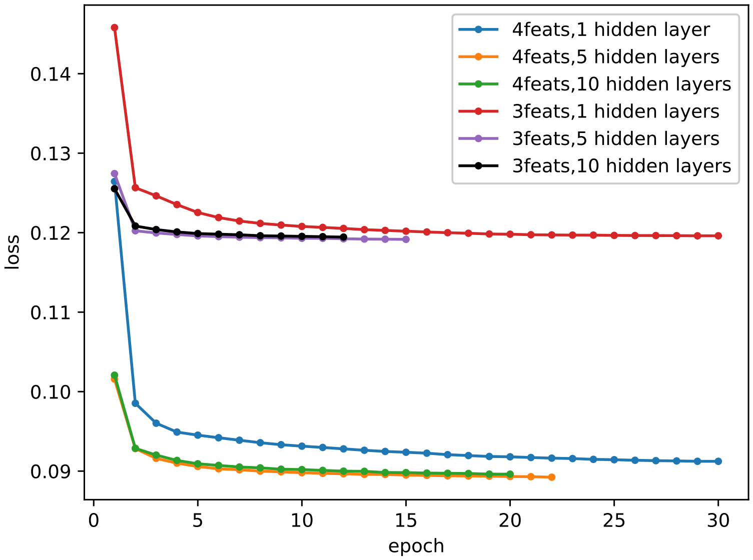


Figure 4.5: The training loss of MLPs trained with three and four features and different hidden layers. The training process of MLP with three features and 10 hidden layers has converged in fewer epochs (12) and MLP trained on four features and five hidden layers has the lowest training error.

probabilities is done with the same approach as Komarov [76]. If this probability is above 95%, the samples is labeled as ice and if the probability is less than 5%, it is labeled as water. If the model predicted probability for a sample is between 5% and 95%, this sample is labeled as unknown. This was done to ensure only samples with high ice and water probability will be used in future data assimilation applications.

Figure 4.5 represents the training loss of all MLP models trained on pure ice and water samples. The figures shows that except for the MLP models with one hidden layer, the other MLP models stopped training before reaching epoch 30. Among these MLP models MLP trained with three features and 10 hidden layers converged in fewer epochs (12) and the MLP trained on four features and five hidden layers has the lowest training error. Table 4.3 displays the accuracy of the discussed models along with their training time and the accuracy of the logistic regression method[76]. The table shows that having the same number of features, increasing the number of hidden layers in the neural network, does not change the accuracy of the model significantly. However, with increasing of the number of hidden layers, the accuracy of the MLP models trained on three features decreases from 80.51% to 78.25. This accuracy reduction is mainly due to the decrease in water classification accuracy. While the classification accuracy of the ice class in this case has variance of 1%, the water class accuracy decreases by 5% based on the number of hidden layers. In addition, the rate of unknown samples is 0.4% and 2.3% less than three features models with 5 and 10 hidden layers respectively. Despite the superiority of one hidden layer network, its total misclassification rate is 0.03% higher than the other two models due to having higher ice misclassification rates.

In contrast to three feature MLP models, when the number of hidden layers is increased, the accuracy of networks trained with four features increases from 85.86% to a maximum of 86.20%. However, for the four feature MLPs, changing the number of hidden layers makes a difference of less than 1% in classification accuracy of each class. This is the same case for misclassification and unknown rates. However, the idea of separating the NWP and SAR wind speeds is shown to increase the accuracy in terms of all measures, specifically the accuracy of water samples (6-12% improvement) which results in increasing the overall accuracy by ∼5%. In addition, the rate of unknown samples has decreased by almost ∼5%. Recalling the numbers of Table 4.1, ∼5% unknown reduction means about 124,000 samples. Comparing the results of MLP networks trained on four features with the Komarov logistic regression method, the MLP approach significantly increases the accuracy of water samples by ∼22%, leading to total accuracy improvement of about 14%. Moreover, the misclassified rate of MLP method is ∼0.07% less than the logistic regression method and the rate of the unknown samples are less than half.

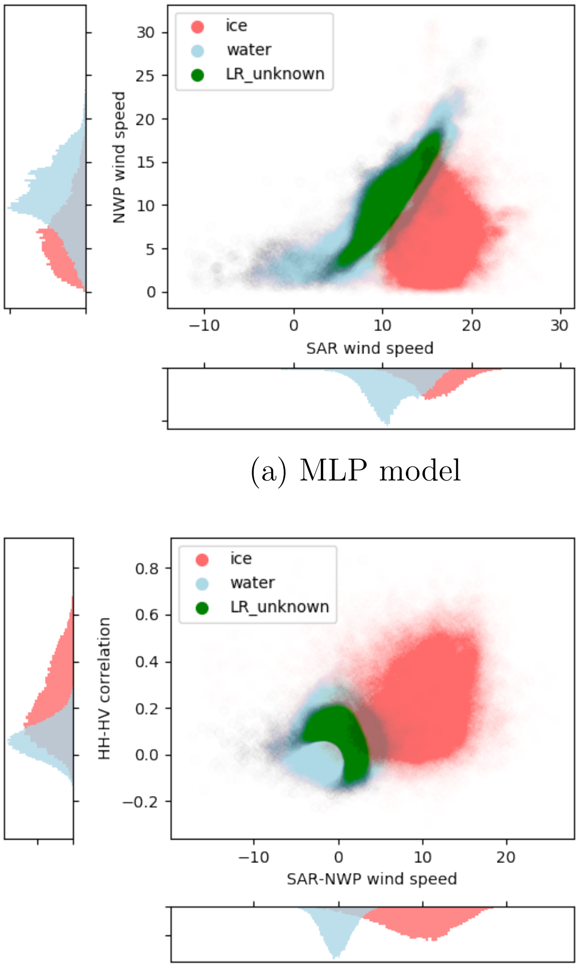
The training time of each MLP model on NVIDIA Titan X GPU is also reported in

Table 4.3: Accuracy of the models trained and tested on pure ice and water samples. MLP_ALL refers to MLP models trained with the whole training set.

Method	Number of features	Number of hidden layers	Training time [min]	Ice accuracy [%]	Ice misclassified [%]	Water accuracy [%]	Water misclassified [%]	Total accuracy [%]	Total misclassified [%]	Unknowns [%]
MLP_ALL	3	1	**6.0**	85.37	1.49	77.28	**0.17**	80.51	0.70	18.79
		5	10.2	86.24	1.42	76.21	0.20	80.22	0.69	19.10
		10	27.5	86.22	1.25	72.95	0.20	78.25	0.62	21.13
	4	1	**6.0**	89.05	1.08	83.74	**0.17**	85.86	**0.53**	13.61
		5	15.5	**90.17**	**1.03**	83.03	0.22	85.88	0.54	13.58
		10	45.5	89.40	1.10	**84.07**	0.19	**86.20**	0.55	**13.25**
Logistic regression	3	-	-	88.23	0.98	61.48	0.35	72.14	0.60	27.25

Table 4.3. Regardless of number of input features, networks with 1 hidden layer required only 6 minutes in total to train while the deep networks required 27.5 and 45.5 minutes. This means that the required training time of the deeper networks is 4.5 and 7.5 times more than shallow networks for networks with 3 and 4 input features respectively.

Figure 4.6, shows how having four features in the MLP significantly increases the accuracy of the open water class when the standard deviation of the SAR wind speed is in the range of [1.5,1.55]. The MLP model for this comparison is the shallow network model represented in Table 4.3 based on its training time efficiency and given that is has a total accuracy close to the maximum accuracy reported in Table 4.3. For the comparison, the water samples of the dataset that have been identified as unknown by the logistic regression, which the MLP model was able to classify correctly, are represented by the green color in the scatter plot. The plot shows how the two classes are separable for the green area using these two features while this is not the case for the logistic regression method based on the scatter plot of the SAR-NWP wind speed versus the HH-HV correlation. This is happening while the histogram of the distributions of the SAR-NWP wind speeds that is represented below the x-axis of the bottom plot is illustrating two separable distributions. It should be noted that Komarov [79] did not force the SAR wind speed to be positive in his wind retrieval approach to obtain wider range of SAR wind speed for the classification. The negative SAR wind speeds mostly happen for smooth open water regions with very low backscatter signals. In addition, this range of SAR wind speed's standard deviation is selected according to its distribution represented in Figure 4.7. The peak of the distribution for both classes is in the range of [1.5,1.55].



(a) MLP model

(b) Logistic regression

Figure 4.6: Cross section of an MLP model and the logistic regression method at the SAR wind speed level of [1.5,1.55]. The MLP model is trained on four features with 1 hidden layer (shown in Table 4.3). The MLP plot (top) shows scattering of samples in the space of SAR and NWP wind speeds and the bottom plot shows the scattering in the space of SAR-NWP wind speed and HH-HV correlation. The green area in both figures, denoted as LR_unknown, shows the water samples in the test dataset that have been identified as unknown by the logistic regression and classified with MLP. Histogram of each feature for the whole test dataset is projected on horizontal and vertical axes.

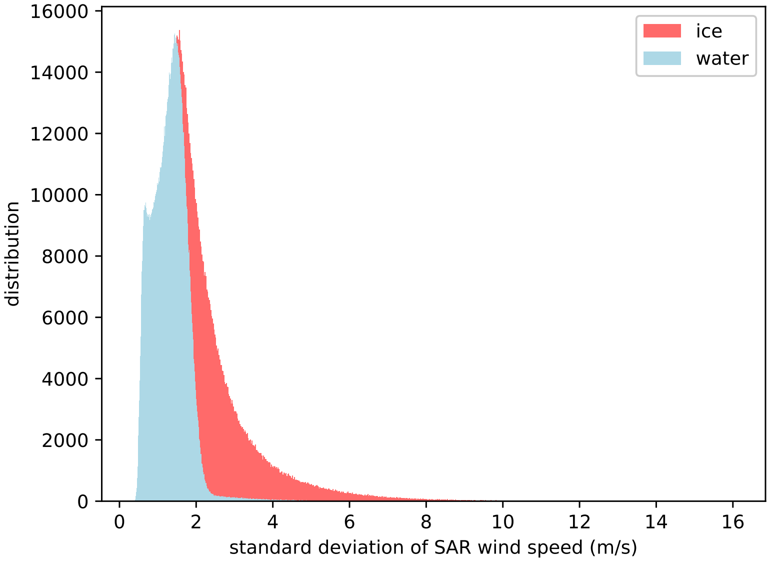
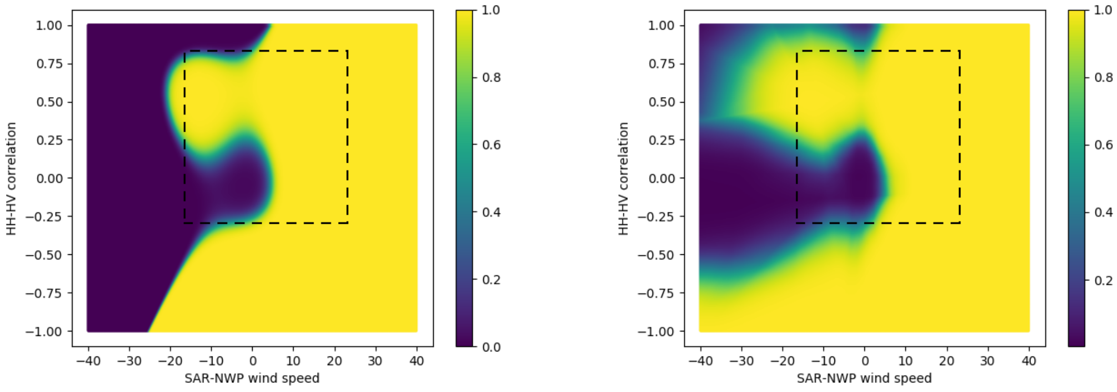


Figure 4.7: Distribution of the standard deviation of SAR wind speed for samples in the test set, which corresponds to data from 2013.



(a) Logistic regression

(b) MLP

Figure 4.8: Probability map of the logistic regression and MLP model when the standard deviation of SAR wind speed is $1.5(m/s)$. The selected MLP model is trained on three features with 1 hidden layer. The black rectangle indicates the training and test area.

Lastly, the probability map of logistic regression is compared to the one corresponding to that from the MLP model trained on three features with one hidden layer. Figure 4.8 represents the cross section of probability maps when the standard deviation of SAR wind speed is $1.5(m/s)$. The figure shows the non-linearity of the MLP decision boundary while the logistic regression's decision boundary is a cubic shape function. Since the bottom left corner of the maps are out of the training and test feature space, there is a disagreement between the compared methods in this region.

4.4.3 Training Using One Year Only

In the next set of experiments, the use of only one year of training data instead of all the six years is investigated. This scenario is more realistic for operational implementation of the method as a long time series of training data is not generally available. The purpose was to see how does the accuracy changes by using the ∼24% of the whole training dataset. The one year subset of all samples is more likely to cover the regions, seasons and weather conditions of the full training dataset compared to drawing random samples from the entire dataset. This context is also more relevant for applications where data may be collected over a previous year, but there is insufficient storage or sensor continuity for longer periods.

The new training set is trained with the same parameters shown in Table 4.2 and Figure

4.9 shows the loss function value over the 30 epochs of an MLP with four features and ten hidden layers. Figure 4.10 represents the training loss of all MLP models trained on 2014 subset of training data. The figures shows that except the MLP models with 5 hidden layers and four features, which has the highest training accuracy, the other MLP models have stopped training before reaching epoch 30. Among these MLP models, the MLP trained with three features and 10 hidden layers has converged in the lowest number of epochs (12).

Table 4.4 shows the numerical results of the experiment evaluated in the same manner as described in Section 4.4.1. While there is no specific relation between the number of hidden layers and changes in the accuracy for networks with specific number of features, the results show that MLP networks trained on four features have higher total accuracy compared to the three feature models. MLP networks with 10 hidden layers have maximum accuracy difference of 6.6% between the four feature and three feature case. In contrast, the three feature MLPs with 5 hidden layers has higher total accuracy and generally, three feature MLPs have much lower water misclassification. Comparing these results with those of Table 4.3, since the training set has become smaller in the current setup, one epoch of the training process will be much faster. As an example, the shallow MLP model would need as low as 25 seconds to train using three features while the same network needs 6 minutes to train on the whole dataset and they both have similar performance. However, the one by one comparison of the MLP networks in Table 4.3 and Table 4.4, shows slightly higher accuracy for the models trained with the whole dataset. This could be because the training set is smaller in this case or because the ice conditions in the two chosen years (2013 and 2014) may differ more substantially from each other than those when the entire dataset is chosen. Comparing the highest accuracy of the model derived with 2014 training dataset with the best model obtained by all training set, the total accuracy of the latter is only about 0.2% higher than the former. Moreover, the model of 2014 dataset has about 0.5% more unknown and 0.07% more misclassified samples. Accordingly, this new model still significantly outperforms the logistic regression model trained on whole dataset as the total accuracy is 14% higher.

4.4.4 Experiment on All Ice Concentrations

The previous section considered MLP performance with respect to ice/water classification for samples corresponding to open water samples and samples with 100% ice concentration. In general, it is desirable to be able to use the MLP to classify an entire SAR image containing a range of ice concentrations. To achieve this goal, Komarov [76] has divided the training set including samples with all ice concentrations into 13 ice concentration

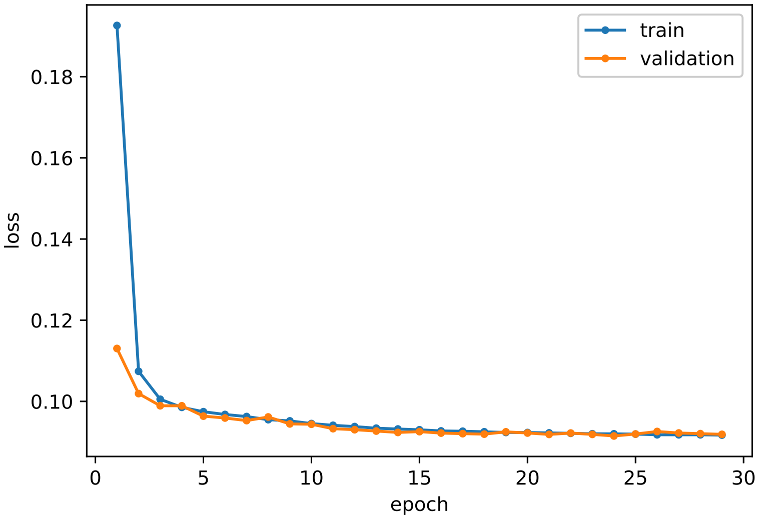


Figure 4.9: The training loss of the MLP trained with four features and 1 hidden layer on the year of 2014 of the training set. The MLP training is stopped after 29 epochs.

Table 4.4: Accuracy of models trained on one year of training set (2014).

Method	Number of features	Number of hidden layers	Training time [min]	Ice accuracy [%]	Ice misclassified [%]	Water accuracy [%]	Water misclassified [%]	Total accuracy [%]	Total misclassified [%]	Unknowns [%]
MLP_2014	3	1	**0.4**	83.29	1.71	78.87	**0.14**	80.63	0.77	18.60
		5	1.3	82.65	1.96	82.37	**0.14**	82.48	0.87	16.65
		10	4.5	83.71	1.58	76.88	0.17	79.61	0.73	19.66
	4	1	1.5	88.33	0.97	80.92	0.19	83.88	**0.50**	15.61
		5	5.0	**91.43**	**0.65**	73.78	0.41	80.83	0.51	18.66
		10	9.0	89.35	1.12	**83.77**	0.29	**86.00**	0.62	**13.77**
Logistic regression	3	-	-	88.23	0.98	61.48	0.35	72.14	0.60	27.25

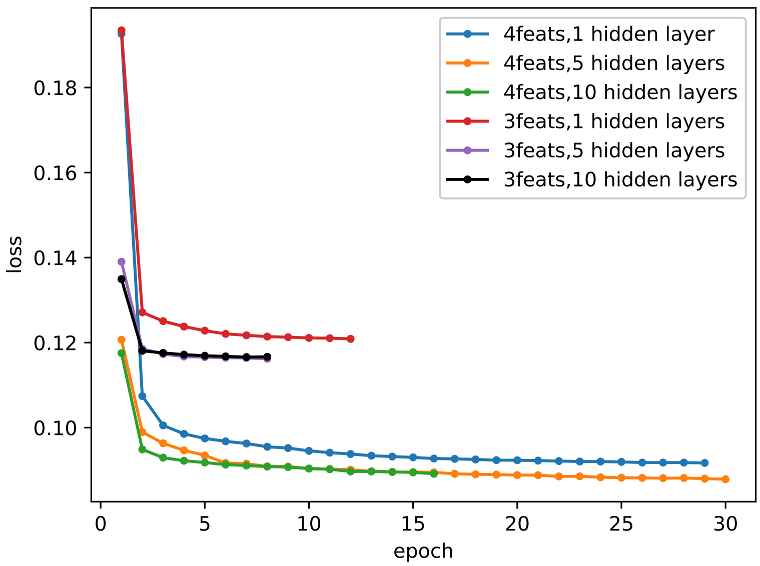
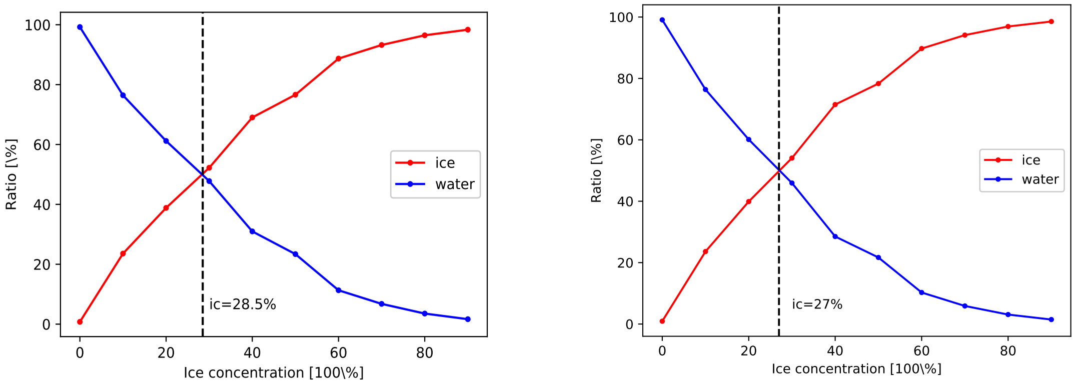


Figure 4.10: The training loss of MLPs trained on 2014 dataset. The training loss of MLP with three features and 5 hidden layers has overlap with the one with 10 hidden layers and converged faster (8 epochs). Moreover, MLP trained on four features and 5 hidden layers has the lowest training error (0.09%).



(a) MLP_ALL

(b) MLP_2014

Figure 4.11: Ratio of ice/water samples from the MLP model predictions for each ice concentration category of CIS image analyses in the test dataset of year 2013. The vertical lines indicate the ice concentration value for the intersection of ratios. MLP_ALL is the MLP model trained on all training data and MLP_2014 is the model trained on samples of year 2014.

categories. These categories are defined based on image analysis products provided by CIS. Afterwards, samples of each ice concentration category were tested with the proposed probability model and the ratio of samples identified as ice to total number of samples identified as ice and water was computed, and the ratio of samples identified as water to the total number of samples identified as ice and water, was also computed. To obtain a threshold that can be used to provide an ice/water label to evaluate the MLP output, the ice concentration value corresponding to the intersection of the two ratios was used.

The approach of obtaining an ice concentration threshold was carried out for test dataset by selecting MLP models with highest accuracy from Tables 4.3 and 4.4 , denoted as MLP_ALL and MLP_2014 respectively, and the result is represented in Figure 4.11. The figure shows that both MLP models have an ice concentration threshold of close to 30%. Since it is ideally expected to have the threshold of 50%, the 30% MLP derived thresholds can be considered more reasonable than 21% from the logistic model. In the next step, the acquired ice concentration threshold is used to evaluate the test dataset with all ice concentrations (of year 2013). The evaluation was performed by the best MLP models (MLP_ALL and MLP_2014) and the results along with Komarov accuracies are reported in Table 4.5.

Table 4.5: Accuracy of the models on all ice concentration test set using their corresponding ice concentration threshold.

method	Number of features	Number of hidden layers	Ice accuracy [%]	Ice misclassified [%]	water accuracy [%]	water misclassified [%]	Total accuracy [%]	Total misclassified [%]	Unknowns [%]
MLP_All	4	10	**81.70**	4.61	**82.54**	2.09	**82.31**	2.77	**14.92**
MLP_2014	4	10	81.37	**4.52**	82.09	2.10	81.89	2.76	15.34
Logistic Regression	3	-	79.41	4.60	65.72	**1.90**	69.65	**2.67**	27.68

Table 4.5 shows that while the accuracy of ice class for all the three models are quite similar, MLP_ALL model's accuracy of water class and its total accuracy is slightly better than MLP_2014 and outperforms the logistic regression method by 17% with 13% less unknown samples. It should be noted that the accuracies in Table 4.5 are lower than model's accuracies reported in Table 4.3, because, the latter shows results of experiments on pure ice and water samples while the former includes all ice concentration categories. This is mainly due to the errors associated with the image analysis for mid-range ice concentrations.

4.5 Summary

In this chapter a new method for the problem of ice/water classification is proposed. This method is developed based on a dataset of 15405 RADARSAT-2 dual-polarization SAR images with their corresponding NWP GEM regional wind speeds which are labeled based on CIS image analysis. This unique dataset includes images from 6 years, different locations, seasons and weather conditions.

The proposed classification method, inspired by the work of Komarov [76], uses 4 different features extracted from window sampling of the SAR images and the interpolated NWP wind speeds. These features are: 1) SAR wind speed, 2) NWP wind speed, 3)HH-HV correlation of the SAR backscatter signals and, 4) standard deviation of SAR wind speed. While Komarov used logistic regression and probability models in his classification method, the proposed approach take advantage of the MLP architecture utilized in popular artificial neural network models. The performance of the model on the pure ice and water samples has been investigated using different hidden layers in the MLP architecture and also training the MLP models with four features and three features. Overall, MLP models accuracy was observed to improve 5-8% when it was trained with four features. The rate of samples classified as unknown also decreased from 19% to 13% using the MLP approach compared to the logistic regression method.

The idea of reducing the number of samples in the training set was also investigated by utilizing only samples of year 2014 to train the MLP models. The total accuracy, at worst case, was only 5% less than the model with similar architecture trained on full training set. Using samples of the year 2014, the MLP with 4 input features and 10 hidden layers outperformed the other models.

Finally, the samples belonging to all ice concentration categories were used for the classification. The threshold of 30% ice concentration was obtained by applying Komarov's approach on the training set on the test set with all ice concentration categories (of year 2013) and the results were compared to his work. The overall accuracy of MLP models was ~12% higher than the logistic regression method and the MLPs were able to have ~12% fewer unknown classified samples with a lower overall misclassification rate. It should be noted that the performance of MLP model trained on 2014 training set was similar to the MLP model trained on full dataset.

Chapter 5

Uncertainty Estimation Using Neural Networks

5.1 Introduction

This chapter is dedicated to illustrate and discuss uncertainty of neural network predictions for the ice/water classification problem using the recent proposed approaches. The modifications to estimate uncertainties are applied to selection of models from Chapter 4. The experiments investigate the relationship between the estimated uncertainties and input features as well as predicted probabilities. The impact of training the neural networks with features of one year or subset of a year on the probabilities and uncertainties is also evaluated in this chapter.

5.2 Uncertainty Estimation

A model is considered to be well-defined if it can make predictions about unobserved data having been trained on observed data. However, any sensible model will be uncertain when predicting unobserved data [46]. Estimating uncertainty of the machine learning models is an often neglected, yet important, task to improve safety and trustworthiness of the developed systems in order to be used practically in real world applications [3]. With recent advances in deep learning, models should be able to predict their uncertainties in addition to their regular predictions. This is also very useful in the field of remote sensing as it helps both data producers and end users to find error characteristics for further improvements in data production and rational use of the data. The uncertainty estimation approaches discussed in the deep learning community mainly consist of two types of uncertainties[29]: 1)*epistemic* uncertainty that is related to uncertainty in the model parameters, and 2)*aleatoric* uncertainty which is due to genuine stochasticity in the data that results in variability in the outcome of experiments [43]. The epistemic uncertainty is considered as either the variance of the model output for a regression problem or the entropy of the probability distribution of the model output for a classification problem. In the following, the approaches from the literature [73, 74] to estimate each of these uncertainties are described. Since the epistemic uncertainty estimation method is motivated by the introduction of dropout in the neural networks, the dropout method is described first.

5.2.1 Dropout Neural Network

Dropout is a popular approach in deep learning applications to reduce overfitting introduced by Hinton et al. [61], and extended by Srivastava [121]. A dropout model is constructed as follows. For a given neural network with L hidden layers, for any layer $0 \leq l \leq L$ of the dropout model with neurons $\mathbf{a}^l$, a binary vector $\mathbf{r}^l$ of the same size is generated by sampling from a Bernoulli distribution with probability $1-p, 0 \leq p \leq 1$. The output of layer l is multiplied element-wise with vector $\mathbf{r}^l$ as

$$\begin{aligned} \mathbf{r}_i^l &\sim \text{Bernoulli}(1-p), \\ \hat{\mathbf{a}}^l &= \mathbf{a}^l \odot \mathbf{r}^l, \end{aligned} \tag{5.1}$$

where $\mathbf{r}_i^l$ is the $i-$th element of vector $\mathbf{r}^l$. In this approach, p portion of outputs of layer l are set to 0, which results in thinned outputs at layer l. This thinned output provides the input for the next layer of the network. Using this approach for each given training case, a sub-network is sampled from a larger network and the back-propagation updates weights of the sub-network. During the test time, the sub-networks are combined by scaling the weights of layer l as $W_{\text{test}}^l = pW^l$ and the dropout layer is ignored.

5.2.2 Epistemic Uncertainty

In theory, epistemic uncertainty due to the model parameters (here, weights) can be modeled by placing a prior distribution over the model's weights and observing the change in the network predictions given the data. Gal and Ghahramani have shown that adding dropout after each layer of the neural network can be interpreted as approximate Bernoulli variational inference to infer the distribution of the weights in Bayesian neural networks [42].

For classification problems, this can be implemented by maintaining the dropout layers at test time, instead of ignoring the dropout layer, and evaluating the model prediction by approximating the predictive posterior using Monte Carlo dropout. Afterwards, the stochastic forward passes through the model are averaged. For a sample $\mathbf{x}_i$ $(1 \leq i \leq N)$ of dataset $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ with size N and labels $\mathbf{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_N\}$ in a binary classification problem, the model prediction is evaluated as

$$\mathbf{p}_i \approx \frac{1}{T} \sum_{t=1}^{T} \text{sigmoid}(\mathbf{f}_i^{\hat{W}_t}). \tag{5.2}$$

In equation (5.2), T is the number of passes, $\hat{W}_t$ is a sample of model weights, $\mathbf{f}_i^{\hat{W}_t}$ represents the output of model for input $\mathbf{x}_i$ and $\mathbf{p}_i$ is the predicted probability. Then, the entropy of the predicted probability can be used to calculate uncertainty of the weights as

$$H(\mathbf{p}_i) = -(\mathbf{p}_i \log \mathbf{p}_i + (1 - \mathbf{p}_i) \log(1 - \mathbf{p}_i)), \tag{5.3}$$

where $\mathbf{p}_i$ and $1 - \mathbf{p}_i$ correspond to ice and water probabilities, respectively. Figure 5.1 shows the entropy value for probabilities between 0 and 1. The figure shows that the entropy is minimum when the probability is either close to 0 or 1, and it is maximum when the probability is 0.5 which means the classifier predicts equal probability for each class. When the distribution of classes in the features space is bimodal, having more data is expected to reduce epistemic uncertainty [73].

5.2.3 Aleatoric Uncertainty

While epistemic uncertainty places a prior distribution over the networks weights to capture the uncertainty in these weights, aleatoric uncertainty is modeled by placing a distribution over the output of the model to capture uncertainty induced by having noisy data, which results in similar features with multiple target labels. Unlike epistemic uncertainty, aleatoric uncertainty does not decrease by additional data since it is assumed to be caused by limitations in the dataset. There are two types of aleatoric uncertainty defined: task-dependent or homoscedastic uncertainty and data-dependent or heteroscedastic uncertainty. Homoscedastic uncertainty assumes constant observation noise for different inputs of a problem while heteroscedastic uncertainty assumes a different value of the noise for each input of the problem. The value of the noise for this case is an additional model output, and learning this value is carried out in an unsupervised manner, through the dependence of the predicted probability on the noise, and the propagation of this dependence through the derivative of the loss function. Heteroscedastic uncertainty is very useful

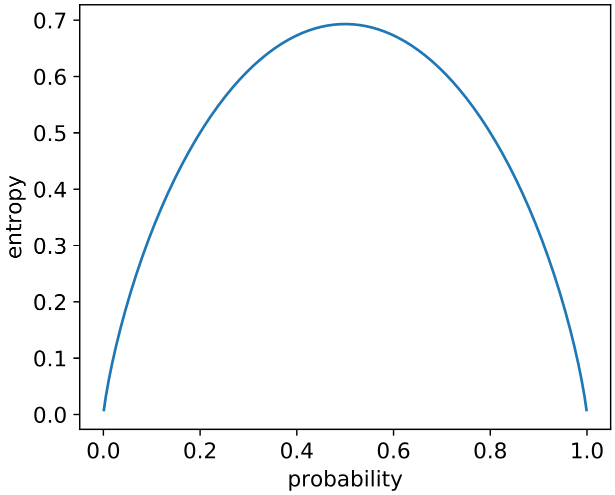


Figure 5.1: Entropy value for different probability values in a binary classification problem. The maximum of entropy is at $p = 0.5$.

in remote sensing applications when data dependent observation noise is present due to measurement conditions and numerical retrieval methods. In Bayesian neural networks [28, 96, 101], this uncertainty can be learned by changing the loss function.

Heteroscedastic uncertainty for regression problems is attained by placing a Gaussian (or Laplacian) prior over the network outputs and predicting the mean and variance of the distribution for each given sample [86]. A similar approach is employed to calculate homoscedastic uncertainty but the estimated variance is a single parameter that shows the task-dependent noise [73].

For classification problems, Kendall and Gal proposed an approach similar to that used for regression to calculate the heteroscedastic uncertainty [73, 74]. For a binary classification problem, the network predicts a unary $\mathbf{f}_i$ for each input i and then this unary is passed through a sigmoid function to produce the probability $\mathbf{p}_i$. The proposed method for calculation of uncertainty places a Gaussian distribution over the unaries where the distribution parameters (mean and variance) are predicted as the model outputs:

$$\begin{aligned} \mathbf{f}'_i | W &\sim \mathcal{N}(\mathbf{f}_i^W, (\sigma_i^W)^2), \\ \hat{\mathbf{p}}_i &= \mathrm{sigmoid}(\mathbf{f}'_i). \end{aligned} \tag{5.4}$$

In equation (5.4), W represents the network parameters and $\mathbf{f}_i^W$ and σ_i^W are the network outputs. For each sample i, output $\mathbf{f}_i^W$ is perturbed by Gaussian noise with variance $(\sigma_i^W)^2)$

and the resulting vector is passed through the sigmoid function to obtain probability vector $\hat{\mathbf{p}}_i$. In this case, the Monte Carlo approach can be used to define the total loss as

$$\mathbf{f}'_{i,t} = \mathbf{f}_i^W + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, (\sigma_i^W)^2)$$
$$L(\mathbf{p}, \hat{\mathbf{p}}) = \sum_{i=1}^{n} \frac{1}{T} \sum_{t=1}^{T} [\mathbf{p}_i \log \hat{\mathbf{p}}_{i,t} + (1 - \mathbf{p}_i) \log (1 - \hat{\mathbf{p}}_{i,t})], \tag{5.5}$$

where $L(\mathbf{p}, \hat{\mathbf{p}})$ is the loss function, n is the total number of inputs indexed by i with label $\mathbf{p}_i$, and T is the total number of Monte Carlo runs for Gaussian sampling indexed by t. The loss function in equation 5.5 ideally would be close to zero when the model predicts low noise variance σ_i and the sigmoid output $\mathbf{f}_i$ yields its either limits (zero or one) corresponding to the true classes. If the model predicts a wrong class because of its logit value, then the variance will be increased to direct the logit value to the opposite direction of the sigmoid function.

5.2.4 Combined Neural Network Model

By combining the method of epistemic uncertainty prediction described in Section 5.2.2 together with the heteroscedastic NN in Section 5.2.3, Kendall and Gal developed neural network models for regression and classification that are able to estimate the epistemic and aleatoric uncertainty at the same time [73, 74]. The combined uncertainty model adds a dropout layer after each hidden layer and produces two outputs, which are the predicted probability and variance. In this case, the combined model uses the Monte Carlo approach twice: 1) during the training to estimate the probabilities along with their corresponding logit variances as aleatoric uncertainty and, 2) during the test time to predict the epistemic uncertainty.

The combined model introduced in this subsection is applied to the ice/water database described in Chapter 4 to generate both epistemic and aleatoric uncertainty maps along with the predicted probability.

5.3 Results and Discussions

5.3.1 Experimental Setup

The uncertainty estimation approaches discussed in this chapter are evaluated to assess their ability to generate meaningful uncertainty maps for the ice/water detection problem described in Chapter 4 and their impact on the accuracy of the classification. Therefore, the same dataset described in Section 4.3 is employed here as well. Experiments are divided mainly in three parts: 1) experiments on pure ice and water samples, 2) experiments on seasonal training of year 2014 and, 3) experiments on samples with all ice concentrations. To keep the problem simple, these experiments only used the MLP models of Chapter 4 with 1 hidden layer.

To setup the experiments regarding the epistemic uncertainty, the dropout rate was set to 5%. For big and deep networks the conventional dropout rate is 50%. However, the networks introduced in Chapter 4 are shallow and small. The other parameter of the epistemic uncertainty estimation is the number of forward passes of Monte Carlo (MC) dropout, which was set to 100 for all the experiments here. Similarly, the heteroscedastic aleatoric uncertainty, which will be referred to as aleatoric uncertainty, has also the MC parameter, which is also set to 100. These parameters were set the same for the combined uncertainty method.

Since the uncertainty estimation approaches required more epochs to converge, the maximum number of training epochs was increased from 30 to 50. The early stopping criteria and other parameters were kept the same. Also, all the variance values were initialized to zero.

5.3.2 Results on Pure Ice and Water Samples

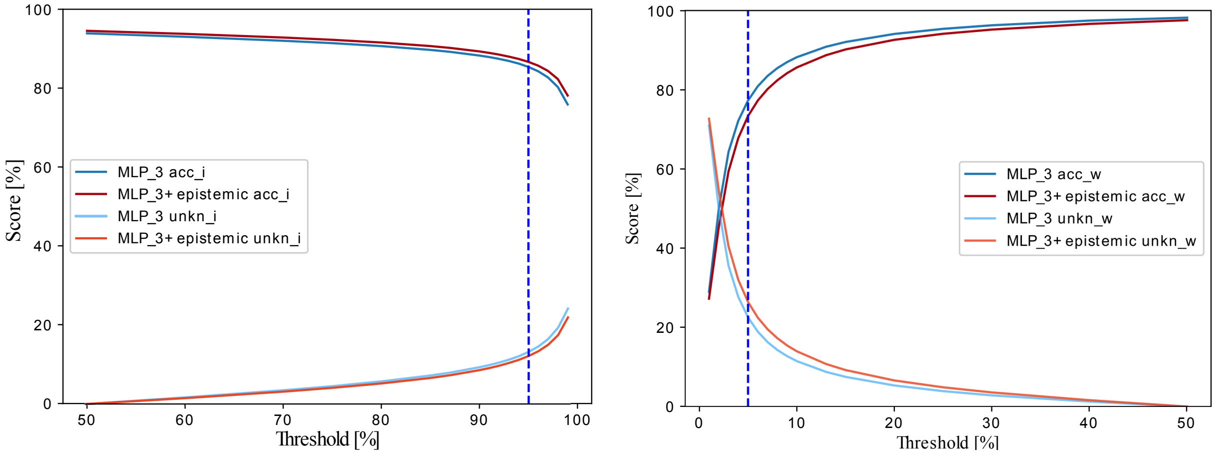
In the following, the results of epistemic uncertainty, aleatoric uncertainty and combination of these two will be represented and discussed.

Epistemic uncertainty

As discussed earlier, the epistemic uncertainty can provide estimates of the uncertainty of the classification output due to the uncertainty of the model weights. Table 5.1 represents the result of experiments on pure ice and water samples when the two types of uncertainties are taken into account. The results of adding epistemic uncertainty to each MLP show overall accuracy reduction by 1-2%. This is mainly due to the network size. Since the number of weights to be trained is significantly less than the number of training samples, adding further regularization by applying dropout will reduce their performance. Likewise, increasing dropout ratio was observed to further decrease the overall accuracy. The classification scores illustrate that some samples of water class were identified as unknown samples, while the accuracy of ice class had small changes. However, the fraction of

Table 5.1: Accuracy of the models trained and tested on pure ice and water samples. Modeling uncertainty in the MLP models gives a slight degradation in total accuracy while notably decreasing the percentage of misclassified samples. The numbers for models with epistemic uncertainty and combined model are average of Monte Carlo runs.

Method	Number of features	Number of hidden layers	Ice accuracy [%]	Ice misclassified [%]	Water accuracy [%]	Water misclassified [%]	Total accuracy [%]	Total misclassified [%]	Unknowns [%]
MLP_3	3	1	85.37	1.49	77.28	0.17	80.51	0.70	18.79
+Epistemic uncertainty			86.66	1.24	73.36	0.23	78.67	0.63	20.69
+Aleatoric uncertainty			86.80	1.20	72.65	0.22	78.30	0.61	21.09
+Epistemic & Aleatoric uncertainty			85.88	1.38	75.89	0.20	79.88	0.67	19.51
MLP_4	4	1	89.05	1.08	**83.74**	0.17	**85.86**	0.53	**13.61**
+Epistemic uncertainty			88.83	0.92	81.20	0.17	84.25	0.47	15.32
+Aleatoric uncertainty			89.64	0.90	80.73	0.19	84.29	0.47	15.24
+Epistemic & aleatoric uncertainty			88.73	0.92	80.96	0.17	84.06	0.47	15.48
MLP_2014	4	1	88.33	0.97	80.92	0.19	83.88	0.50	15.61
+Epistemic uncertainty			88.16	0.82	78.55	**0.15**	82.39	0.42	17.18
+Aleatoric uncertainty			**89.65**	0.77	76.94	0.24	82.02	0.45	17.53
+Epistemic & aleatoric uncertainty			89.21	**0.58**	72.47	0.20	79.15	**0.36**	20.49
Logistic Regression	3	-	88.23	0.98	61.48	0.35	77.16	0.60	27.24



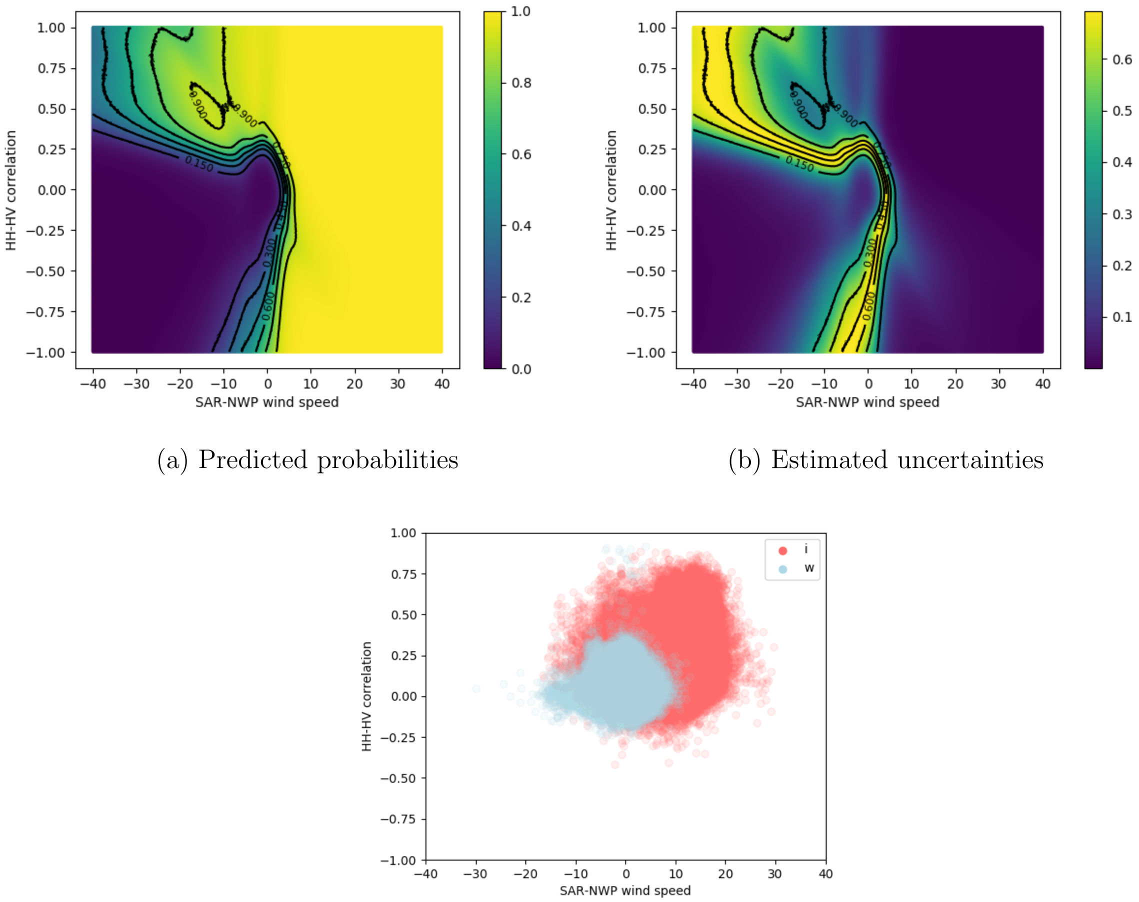
(a) Ice class

(b) Water class

Figure 5.2: The impact of probability thresholds used to define ice and water points on the accuracy of ice and water class. The selected models are MLP with three features with and without epistemic uncertainty. The 5% and 95% probability threshold for ice and water is shown by a vertical line in each case.

misclassified ice samples decreases in all cases and for MLP_2014 the fraction of both misclassified ice and water decreases. This relationship between the accuracies and unknown ratios is represented in more detail by Figure 5.2 where the accuracy of ice and water class as well as their unknown percentage is plotted versus a variation of thresholds for MLP_3 with and without epistemic uncertainty. The figure shows that adding the epistemic uncertainty changes the scores of water class more than ice class. In addition, the figure illustrates that the choice of threshold has more influence on the water class rather than the ice class as for the water thresholds below 2%, the majority of samples will be labeled as unknown rather than correct water label. This means that the selected classification models have more power in predicting close to 100% probability for the ice samples in the test dataset rather than predicting 0% probability for the water samples.

Figure 5.3 visually represents the epistemic uncertainty and probability maps of the results for the MLP_3 model when standard deviation of SAR wind speed is 1.5. The figure also shows the HH-HV correlation and SAR-NWP wind speed scatter in the training dataset when the std of SAR wind speed is 1.5± 0.02. As expected, the model uncertainties are greater on the decision boundary and its extension into unobserved spaces of the feature space.



(a) Predicted probabilities

(b) Estimated uncertainties

(c) Scatter plot of the training features

Figure 5.3: The probability (a) and model uncertainty (b) map of the MLP_3 model trained on three features with epistemic uncertainty when the standard deviation of SAR wind speed is 1.5. The scattering of the training dataset features when the std of SAR wind speed is in the interval of [1.48-1.52] is represented in (c). The contour lines of predicted probabilities are overlaid on panel (a) and (b). In panel (c) red dots indicate ice points while grey dots indicate water. The white region is the unobserved region. The model uncertainty is higher on the decision boundary and where this boundary extends into unobserved regions.

Aleatoric uncertainty

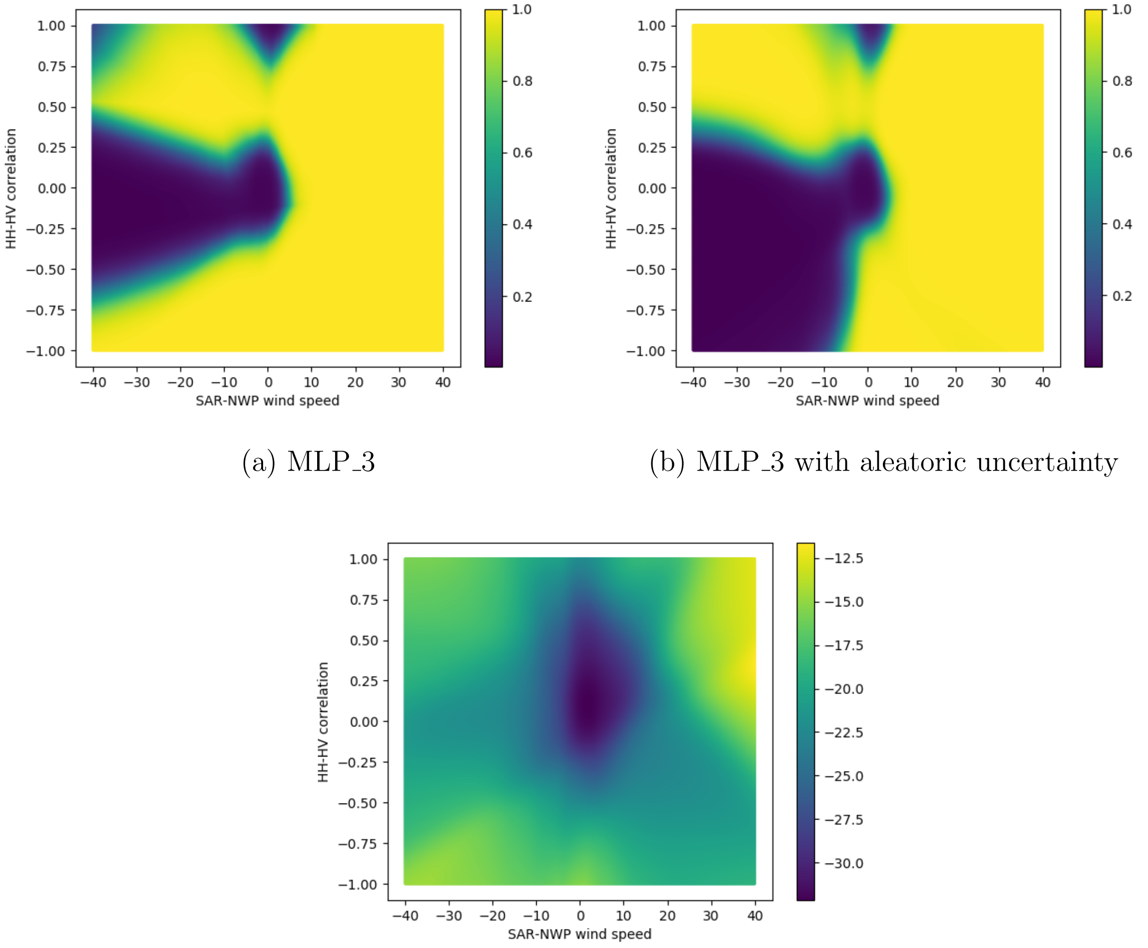
The results of employing aleatoric uncertainty in the MLP models are also shown in Table 5.1. Similar to epistemic uncertainty, adding aleatoric uncertainty slightly reduces the water accuracy and total accuracy of the MLP models. The total accuracy for each MLP model in this case is decreased by 2-4%. In addition, adding uncertainty in this case increases the unknown ratio similar to the epistemic uncertainty. Despite these accuracy reductions, for each MLP category and its variations of added uncertainty, the models with added aleatoric uncertainty have lower ice misclassification rate and higher ice accuracy and their overall accuracies are higher than those of logistic regression. The decision boundaries of this model and MLP_3 model is illustrated in Figure 5.4. Figure 5.5 also represents the distribution of the logits and their predicted variances for MLP_3 with added aleatoric uncertainty. The vertical lines in panel (a) of the figure show how the misclassification rate of ice class is higher than water class because the distribution for the ice samples crosses into the region with probabilities less than 0.05, corresponding to water.

Combined uncertainty

The combined model takes the advantage of obtaining both model's uncertainty as well as input's uncertainty at the same time. As expected, the accuracies reported in Table 5.1 show that the combined model modifications on the MLPs does not increase their total accuracy and the results are similar to having only either aleatoric or epistemic uncertainty. However, the ice and total misclassification rate of MLP_2014 is minimum with the score of 0.58% and 0.36% respectively. These scores are almost half of logistic regression misclassification scores.

5.3.3 Results on Subsets of a Year

In this part of the experiments, the training samples from the year of 2014 were divided into 3 subsets based on their month of acquisition. For each subset, an MLP was trained to see if it is possible to have reasonable outcomes if for any reason, we only have data from a specific period of the year available for training. In addition, we wanted to see how the epistemic and aleatoric uncertainties are able to meaningfully represent these limitations. For simplicity, all the models are trained with 1 hidden layer and three features. The MLP models are trained on subset of: 1) January to April (MLP_JA), 2) May to August (MLP_MA),and 3) September to December (MLP_SD). The distribution of training samples for each subset along with the entire year of 2014 and the test dataset is displayed

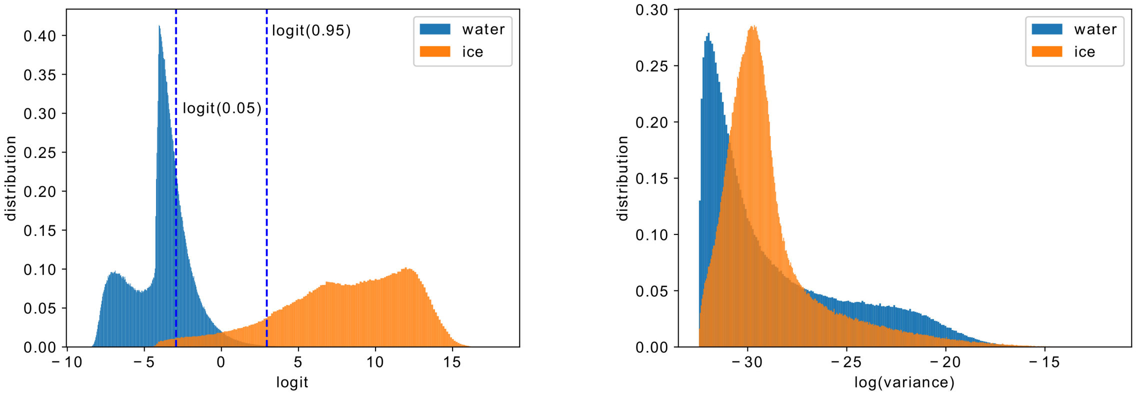


(a) MLP_3

(b) MLP_3 with aleatoric uncertainty

(c) Estimated uncertainty

Figure 5.4: The probability map of the MLP model trained on three features without aleatoric uncertainty (a), and with aleatoric uncertainty (b) when the standard deviation of SAR wind speed is 1.5. The aleatoric uncertainty map of the latter is shown in (c) (logarithmic scale). The aleatoric uncertainty is in direct relation to the distance from the center of feature spaces.



(a) predicted logits

(b) predicted variance

Figure 5.5: The distribution of predicted logits (a) and variances (b) for the MLP_3 model trained with aleatoric uncertainty. The classification boundary of each class is shown by a vertical line in panel (a).

in Figure 5.6. The histograms show that except for the period of May to August, the remaining training data sets have similar distribution regarding each feature. Table 5.2 also shows the size of each subset and their results on the prediction of the test dataset.

The scores in Table 5.2 show that the portion of January to April (MLP_JA) is able to correctly classify ice samples better than other models with the highest score of accuracy and lowest score of misclassification. However, the model trained based on these data performs poorly on water samples as the water class accuracy is below 50%. In contrast, MLP_MA which is trained mostly on summer data, has the highest ice misclassifications, which leads to the highest total misclassification scores. This is predictable based on the PDFs displayed in Figure 5.6 where it shows the PDF of test dataset and other models are different from MLP_MA model is trained on different PDFs.

5.3.4 Impact of Each Feature on Results

To investigate the impact of each feature on the predicted probabilities and uncertainties, Table 5.3 and 5.4 are provided. Table 5.3 is based on predictions of the MLP_2014 and MLP_MA models, trained with epistemic uncertainty, on the test dataset. The main reason of selecting MLP_MA to compare with MLP_2014 is its different training distribution and limited number of samples as shown in Figure 5.6 and Table 5.2. Table 5.3 shows the

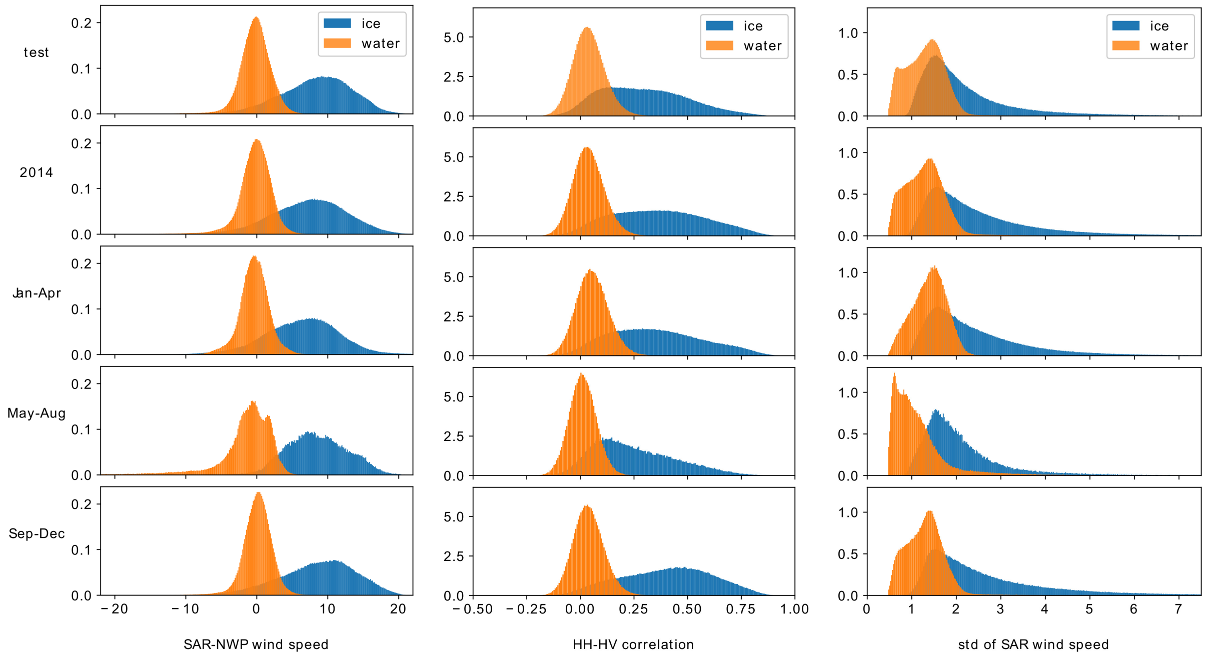


Figure 5.6: Distribution of ice and water samples for the 3 input features of the test dataset, the entire year of 2014, and its three subsets.

Table 5.2: Accuracy of the models trained on the year of 2014 and its 3 subsets each covering 4 months. The models are tested on all pure ice and water samples.

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

predicted probabilities and epistemic uncertainty on the distribution of ice and water class for each feature. The first two columns in each row represents the probability, while the next two represent the uncertainty. Table 5.4 is provided in a similar way but using MLP_2014 and MLP_MA models trained with aleatoric uncertainty so the uncertainty colorbars are representing the logarithm of aleatoric uncertainty.

Table 5.3 shows that SAR-NWP wind speed is the feature causing the most misclassifications as it has more green color on the left side of the ice distribution indicating a low probability of ice. The number of ice misclassifications appears greater for MLP_MA as can be seen from the enhanced green region, consistent with Table 5.2. In addition, Table 5.3 shows how having the big overlap in the distribution of SAR wind speed's standard deviation in the training and test data, as shown in the last column of Figure 5.6, is causing a wider region of high uncertainty. This uncertainty is higher for MLP_2014 comparing to MLP_MA since the ice and water distributions of the training set in the MLP_MA are more separable. Table 5.4 implies similar conclusions about the predicted probabilities of the models trained with aleatoric uncertainty. However, the uncertainty color of MLP_MA PDFs illustrate that uncertainty of both classes in this model are much higher than MLP_2014, especially for the third feature. Additionally, Table 5.3 shows that the aleatoric uncertainty is higher on the left side of ice distribution and right side of water distribution in each case since these regions are close to unobserved space in the feature space.

5.3.5 Results on All Ice Concentration

Table 5.5 shows impact of adding uncertainty to the MLP models when samples covering all ice concentration values are used for testing. The predicted probabilities of MLP models were converted to 0 and 1 labels using a 30% threshold obtained in Chapter 4. Similar to the results on pure ice and water samples, adding uncertainty was observed to reduce the overall accuracy of their original MLP models by about 2% but the misclassified rate was also reduced in most cases with the cost of increasing the unknown labels in all cases. Generally, the water class accuracy and overall accuracy of the MLP_4 was better than other models with scores of 82.54% and 82.31% respectively.

While the misclassification rates for water samples have low variation from 1.30% (MLP_2014) to 2.38% (MLP_4 with aleatoric uncertainty), the misclassification scores for ice class are more variable. The maximum score is 5.58% for MLP_2014 and minimum score is 3.38% for MLP_2014 when both uncertainties are added. The latter also has the minimum overall misclassification score, 2.37%.

Table 5.3: Impact of each feature on the misclassification and epistemic uncertainty of the MLP_2014 and MLP_MA trained on three features.

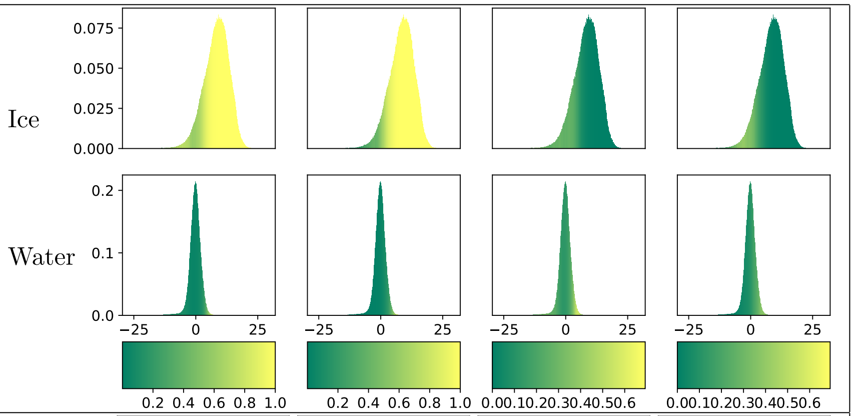
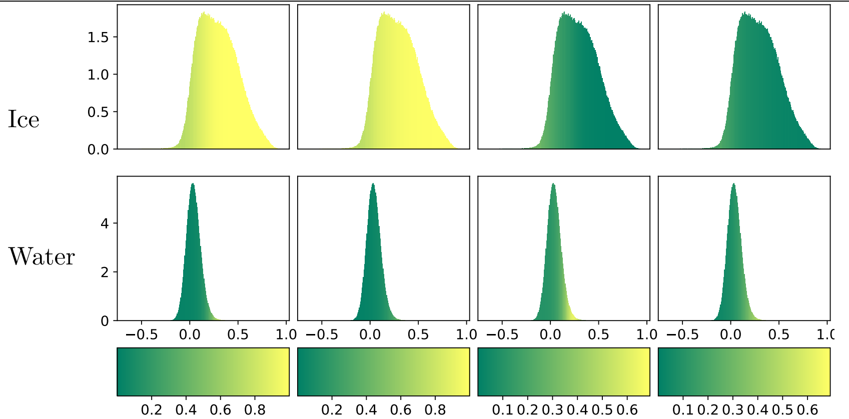
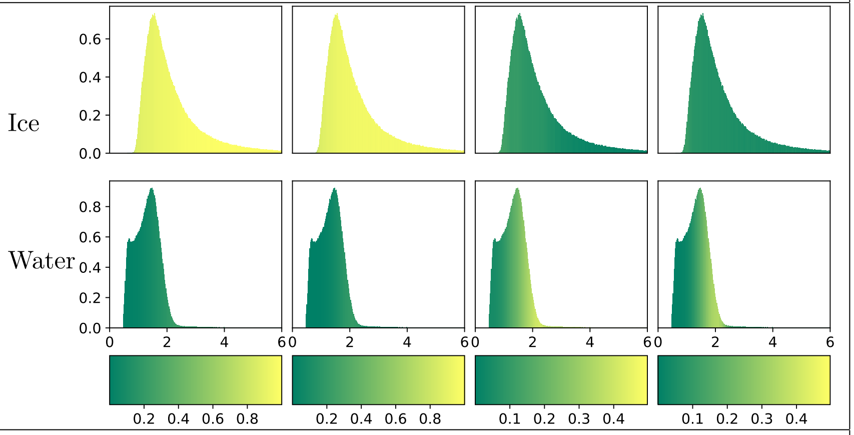
SAR-NWP wind speed	
HH-HV correlation	
std of SAR wind speed	
Model	MLP_2014 MLP_MA MLP_2014 MLP_MA
Colorbar	Predicted probability Epistemic uncertainty

Table 5.4: Impact of each feature on the misclassification and aleatoric uncertainty of the MLP_2014 and MLP_MA trained on three features.

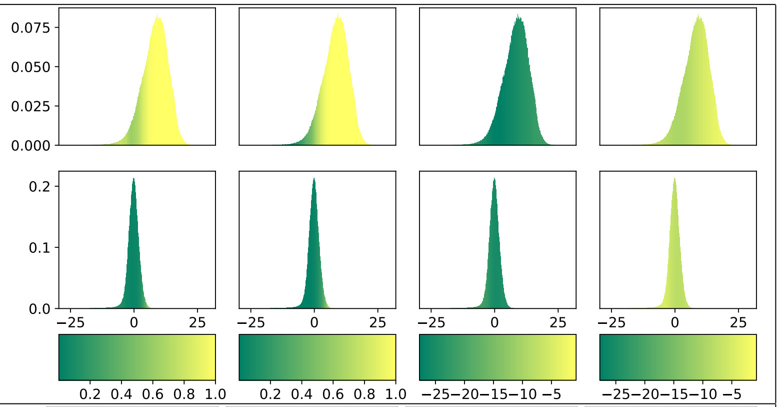
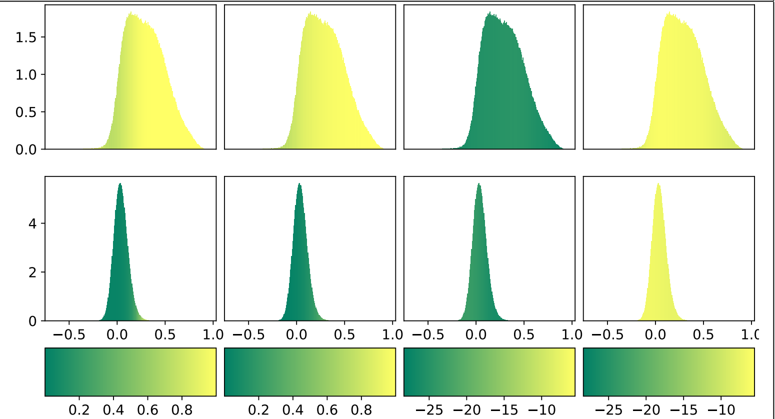
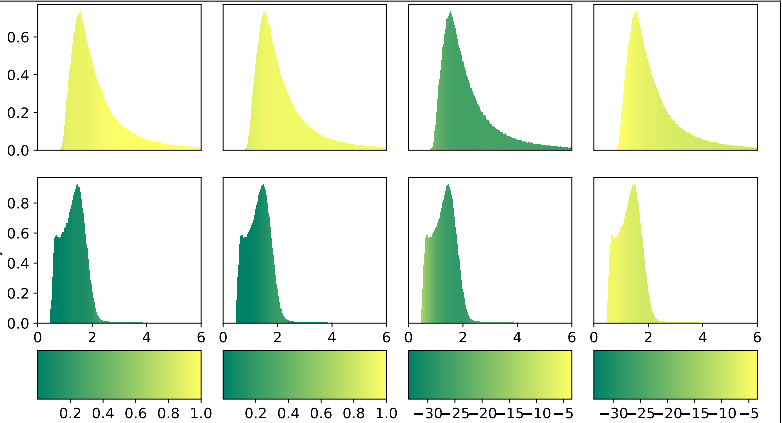
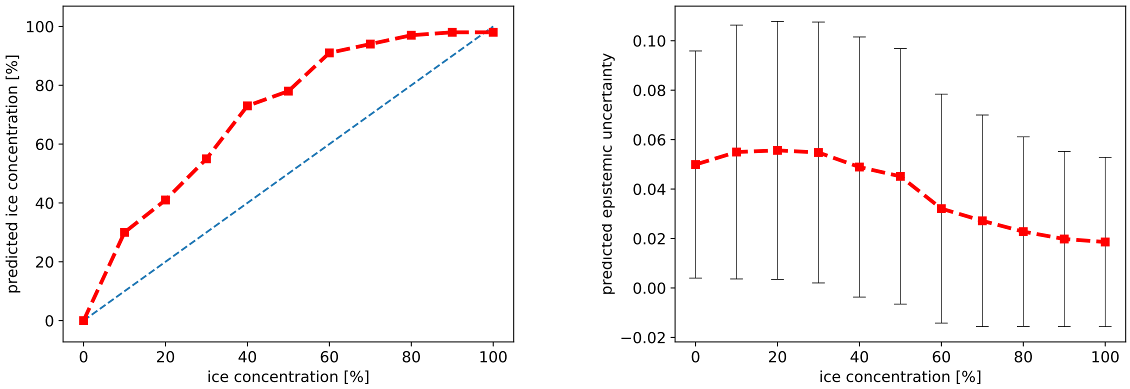
SAR-NWP wind speed	Ice Water 
HH-HV correlation	Ice Water 
std of SAR wind speed	Ice Water 
Model	MLP_2014 MLP_MA MLP_2014 MLP_MA
Colorbar	Predicted probability Aleatoric uncertainty

Table 5.5: Accuracy of the models on all ice concentration test set using the 30% ice concentration threshold. The numbers for models with epistemic uncertainty and combined model are average of Monte Carlo runs.

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



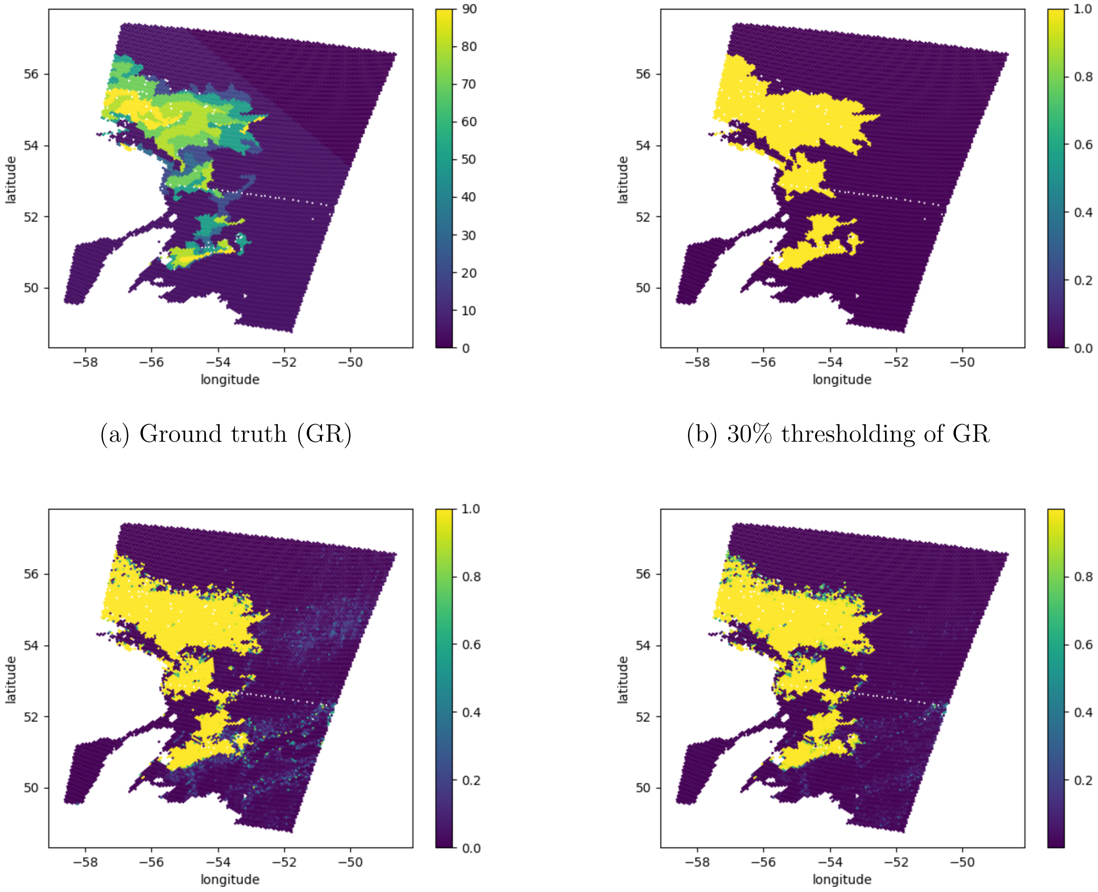
Ice concentration

Epistemic Uncertainty

Figure 5.7: MLP_4 predicted ice concentration (a) and epistemic uncertainty (b) for each ice concentration category. The blue line in (a) shows where the predictions are expected to happen. Vertical lines of panel (b) indicate the standard deviations.

To show how the models predict the ice concentration according to each category, Figure 5.7 is presented for MLP_4 model as an example. For samples of each ice concentration category on the ice chart, the average calculated ice concentration by the MLP_4 model is plotted in Panel (a) and the average estimated uncertainty in Panel (b). The ice concentration in each category is calculated as the ratio of samples predicted as ice to the total ice and water predicted samples in that category. In addition, the vertical bar for each ice chart ice concentration category in panel (b) shows the standard deviation of the estimated uncertainties for samples in the corresponding ice concentration category. Panel (a) of the figure shows that almost for all ice concentrations the model overestimates the ice concentration by about 20%. Panel (b) also shows that the peak of the uncertainty is at 30% ice concentration and its decreasing on both side of this value.

Figure 5.8 represents an example of predictions made by logistic regression and MLP_4 with combined uncertainty for an image acquired on May 3, 2013 over the Labrador Sea. While the logistic regression method has difficulties classifying the water samples correctly, the combined MLP model is able to reduce the misclassified water samples. Moreover, the combined model has the advantage of providing the model uncertainty as well as input induced uncertainty. In addition, MLP model results preserve the high resolution of logistic regression. As represented in Figure 5.9, the model has difficulties in classifying water samples was due to the existence of negative SAR wind speed and high wind speed's



(a) Ground truth (GR)

(b) 30% thresholding of GR

(c) Logistic regression

(d) combined MLP_4

Figure 5.8: An example of ground truth and estimated probabilities for an image acquired on May 3, 2013, over the Labrador Sea.

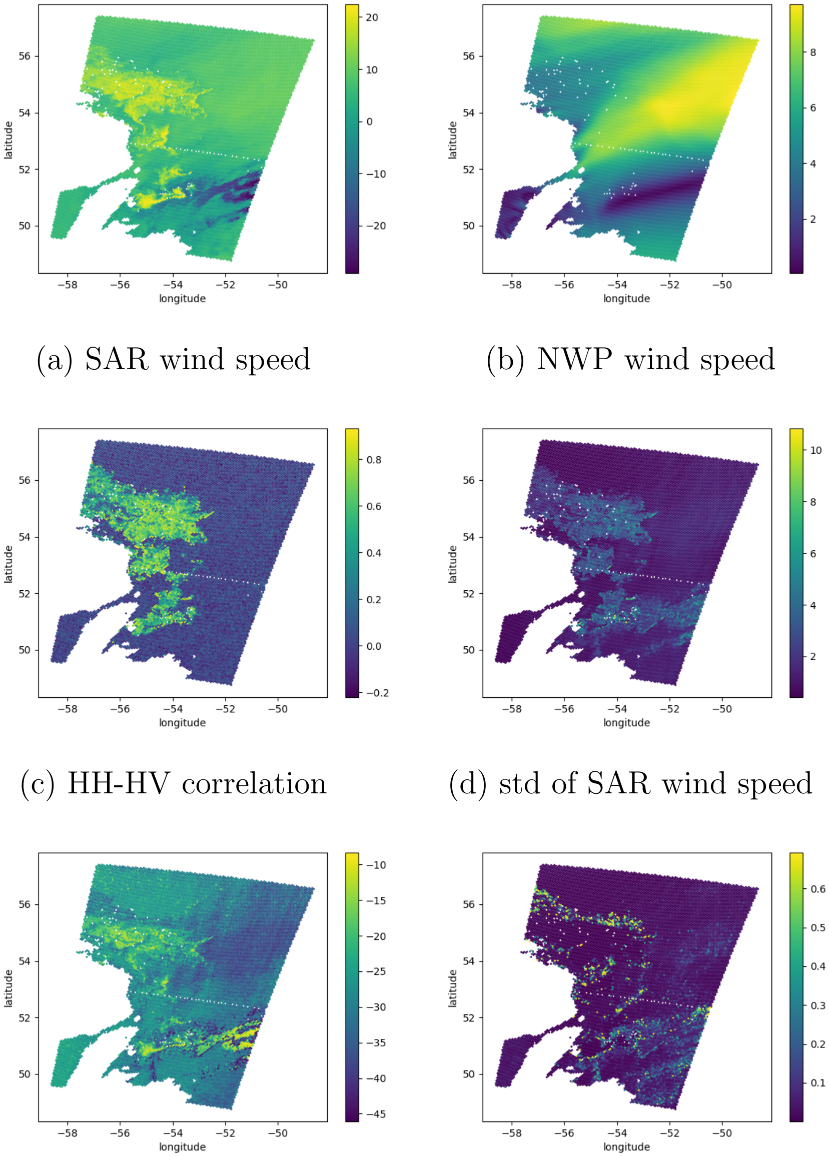
standard deviation over open water areas shown in panels (a) and (d). The uncertainty maps of the MLP_4 combined model are shown in Figure 5.9. The figure shows that the regions with elevated uncertainty are correlated with high wind speed over open water (aleatoric uncertainty) and the uncertainty due to the model (epistemic uncertainty) on the boundary of ice and water can also be seen.

5.3.6 Uncertainty as a Measure of Accuracy

In all the experiments discussed so far, the model uncertainty, epistemic uncertainty, was utilized to visually compare the uncertainty map with the prediction map and ground truth or observe its impact on the classification accuracies. In this part, the correlation between the estimated epistemic uncertainty of an image and its classification accuracy is being investigated. If such a relationship exists, the uncertainty values can be used to estimate the quality of predictions for operational applications without requiring any additional data. The experiments of this section are based on the predictions produced by MLP_3 trained with epistemic uncertainty on the 2013 test dataset containing all ice concentration categories. For each group of samples belonging to a single image of this dataset, the mean of classification accuracy and uncertainty is calculated. Panel (a) of Figure 5.10 shows how the mean accuracy of each image is strongly correlated with the mean epistemic uncertainty of that image. The regressed line on the plot has an R-squared of 0.94 which means the classification accuracy can be calculated based on the epistemic uncertainty with a high precision. The red point of panel (a) corresponds to the SAR image discussed in Figures 5.8 and 5.9 with the mean accuracy of 85.37%. Also, the point with cyan color belongs to a SAR image from East of Baffin Island with mean accuracy of 88.38%. The original ground truth and its ice/water labels after applying a 30% threshold on the ice concentration values are represented in panels (b) and (c) respectively. In addition, panel (d) shows the estimated epistemic uncertainties and panel (d) represents the predicted probabilities for this SAR image.

5.4 Summary

This chapter was dedicated to investigate the new approaches proposed in the neural network community to capture the uncertainties induced by model weights and input features. The epistemic uncertainty which captures the uncertainty of the model due to its weights, adds dropout layers to the model and runs Monte Carlo simulation during test time to calculate the uncertainty. The aleatoric uncertainty is derived by adding another output



(a) SAR wind speed

(b) NWP wind speed

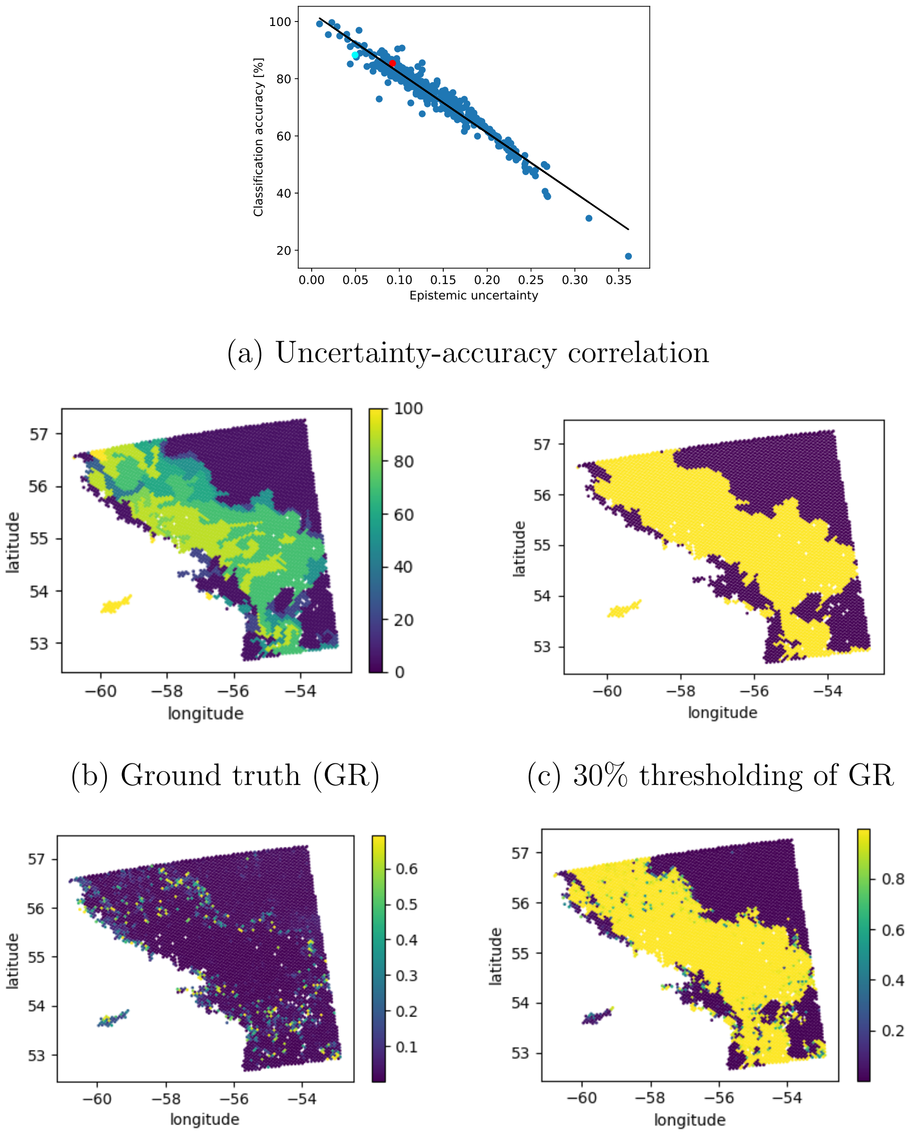
(c) HH-HV correlation

(d) std of SAR wind speed

(e) aleatoric uncertainty

(f) epistemic uncertainty

Figure 5.9: Feature maps and uncertainty maps for the example shown in Figure 5.8 regarding the image acquired on May 3, 2013, over the Labrador Sea. Aleatoric uncertainty is shown in logarithmic scale.



(a) Uncertainty-accuracy correlation

(b) Ground truth (GR)

(c) 30% thresholding of GR

(d) Epistemic uncertainty

(e) predicted probability

Figure 5.10: Correlation between the epistemic uncertainty and accuracy of MLP_3 with epistemic uncertainty (a). Each point of the panel (a) corresponds to average results of samples from a single image of the 2013 dataset. The black line indicates the outcome of linear regression with R-squared of 0.94. The red point corresponds to the SAR image represented in Figure 5.8. The cyan point corresponds to a scene with prediction accuracy of 88.3% represented in panels (b) to (e).

to the network to predict the variance of logits. In this case, for each input 100 samples are drawn from a Gaussian distribution with zero mean and variance corresponding to the current estimate of aleatoric uncertainty and the average loss of these samples is used to update the network parameters during the training. The proposed methods to capture these uncertainties and their combination were tested on the ice/water classification problem discussed in the previous chapter. The results revealed that the added uncertainties reduced the accuracies by 2-3% in all cases. However, the misclassification scores were also reduced. The combined uncertainty models also have the benefit of generating maps of both epistemic and aleatoric uncertainty along with predicted probability maps which can be very useful for practical application. Lastly, the results indicates a strong correlation between the mean accuracy and mean epistemic uncertainty when the mean is calculated using samples from a single image. This indicates that uncertainty may be used to flag individual images with low classification accuracy.

Chapter 6

Conclusion

Motivated by the need to have more accurate predictions of the sea ice cover, this thesis has explored data-driven approaches with the overarching goal of improving the quality of predictions in a sea ice data assimilation system. Two different investigations were carried out. The first examined an alternative cost function designed to enable sharp features to be retained in the analysis of the data assimilation system. The second examined a neural network approach to extract ice/water observations and their uncertainties from SAR sea ice imagery. Even though these approaches can be applied to any sea ice information that meets the specified assumptions, the current research was conducted on real observations of ice thickness (Chapter 3) and ice/water samples (Chapter 4,5). A summary of the thesis contributions is given in the following section followed by a section discussing future work.

6.1 Summary of Contributions

Traditional data assimilation schemes are not usually able to retain sharp features in the observational information. This may be due to spatial averaging of high resolution information or background error correlations, which spread observational information spatially. This latter aspect depends on the details of the background error correlations. In Chapter 3, it is demonstrated, for the first time, that sea ice thickness exhibits a sparse representation in the derivative domain. This has been demonstrated using (1) sea ice thickness measurements from an AEM sensor over the Beaufort Sea; (2) submarine upward looking sonar data; and (3) sea ice thickness derived from Cryosat.

Sharp features in the sea ice thickness states are similar to sharp edges in the image processing problems. When the distribution of the image spatial derivative is sparse, this implies the distribution is sharply peaked at zero and only a small portion of the elements corresponding to edges are non-zero. Chapter 3 introduces and evaluates the use of an additional term in the objective function to constrain sparsity on the derivative of the ice thickness state to retain sharp features when using sea ice thickness data in a data fusion or data assimilation scheme. This l_1-l_2 formulation is compared with the standard l_2 regularization first using data fusion, and then by carrying out data assimilation experiments using a toy sea-ice model. For data fusion, a clear benefit to the l_1-l_2 formulation is observed when the background correlation error length scale is small (on the order of twice the analysis grid spacing). It can be expected that in the vicinity of a sharp feature (e.g., ice edge) the background error correlation length scale may be in this range [17]. This data fusion result could be relevant for the generation of merged sea ice products, where sharp features are desired [7].

For data assimilation, a clear benefit is also attributed to the l_1-l_2 regularisation, although the impact of the error correlation length scales on the difference between the l_1-l_2 method and the l_2 method is less clear. This may be due to the spatial averaging of ice thickness that was required to increase the scale of the data, or it could be due to the model dynamics. However, based on the preliminary results, the l_1-l_2 method is superior with regards to capturing openings in the ice cover than the conventional l_2 method. This was observed for a variety of error correlation length scales, values of the regularisation parameter, and model initial conditions.

In Chapter 4, the problem of providing accurate ice and water observations for data assimilation systems and improving the quality of automated SAR-based ice/water classification, was investigated using a class of neural networks, MLPs. The result of this study was compared with a previous study on the same dataset using logistic regression approach for classification. This study is based on a unique database provided by Environment and Climate Change Canada consisting of four features including SAR retrieved wind speed, NWP wind speed, HH-HV correlation and standard deviation of SAR wind speed, from 15405 RADARSAT-2 HH-HV ScanSAR images with their corresponding CIS Image Analysis information. The database covers the period of time between November 1, 2010 and September 30, 2016. To utilize only samples with high probabilities in future data assimilation applications, only samples with probabilities greater than 95% and less than 5% were labeled as ice and water respectively and the remaining samples were labeled as unknown. The logistic regression approach is developed on three features of this dataset where the difference between SAR and NWP wind speeds are employed as one feature instead of using them separately. The comparison between the three feature version of MLPs with logistic regression results revealed that the MLP models were able to classify water samples with higher accuracy and lower misclassifications and overall reduced rate of unknown labels. However, the total misclassification rate was higher in MLPs. Moreover, increasing the number of hidden layers toward a deep network was observed not to significantly impact the accuracy of the trained models. In all cases, that differences were noted between the two approaches in particular for the water class. The MLP classifiers were able to reduce the ratio of unknown samples by 50% in some cases in comparison to the logistic regression using the same fixed thresholds for probability. It should be noted that Environment and Climate Change Canada is planning to use a developed version of this logistic regression approach where the results are slightly improved [77].

In Chapter 5, the recently proposed uncertainty estimation approaches in the area of neural networks, was employed on the ice/water classification problem of Chapter 4 to produce uncertainty maps in addition to ice/water labels for the first time. The modified MLPs could predict the uncertainty induced by model parameters (weights) as well as noise inherent in the input features. Models were investigated where these two types of uncertainty were considered separately, and also simultaneously in a combined model. The experimental results from the models including uncertainty indicated slightly reduced misclassification rate and increased unknown rates. In this chapter, the impact of each feature on the misclassification rate and uncertainties of each class was also investigated. The visual comparison of the probability and uncertainty maps with the CIS Image Analysis chart implies that the predicted uncertainties may be useful to flag regions in the MLP predictions that should be checked manually by an analyst. The lower misclassification rates that are achieved when uncertainty is included in the model also suggests that these ice/water observations may be useful for data assimilation.

As a summary, the contributions of the dissertation can be listed as:

- Demonstrating the sea ice thickness sparse representation in the derivative domain using data from different observing systems,

- Evaluating the use of l_1-regularization in data fusion and data assimilation experiments with different observation and background error correlation length scales to retain sparsity of sea ice thickness data,

- Improving the ice/water classification accuracy of logistic regression using neural network approaches,

- Providing model and input uncertainty maps with the classification products and investigating their relationship with the input features.

6.2 Future Work

The work presented in this dissertation will provide a foundation for future studies on regularized sea ice data assimilation and uncertainty estimation. This research can be pursued in several possible paths that are presented in the following.

1. Using the l_1-l_2-norm regularization approach to fuse ice thickness observations from multiple sources such as CryoSat and SMOS.

2. Evaluate the l_1-l_2-norm regularization approach on a data assimilation experiment with an operational sea ice model such as Los Alamos sea ice model (CICE) which is currently being used in Environment Canada regional ice-ocean prediction system (RIOPS).

3. Using convolutional neural networks (CNNs) to directly train high resolution SAR images of the ice/water classification problem instead of training MLPs on pixel samples with a tabular format. The recent studies have shown that CNNs can be very useful when there is a relationship between spatial features in the input data and the target labels which is the case for ice and water regions in the SAR images. Moreover, the high resolution CNN outputs can be used to extract ice concentrations.

4. The relationship between the epistemic and aleatoric uncertainties and the classification errors needs additional quantitative investigations. Additionally, it might be interesting to study the link between the uncertainties and quality of the training samples by training the models with all ice concentrations instead of 0% and 100% ice concentration categories, or the link between uncertainties and training size. Use of another measure to quantify epistemic uncertainty instead of entropy can also be explored.

References

[1] H. Akaike. Likelihood and the Bayes procedure. *Trabajos de estadística y de investigación operativa*, 31(1):143–166, 1980.

[2] W. Aldenhoff, C. Heuzé, and L. Eriksson. Comparison of ice/water classification in Fram Strait from C-and L-band SAR imagery. *Annals of Glaciology*, 59(76pt2):112–123, 2018.

[3] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

[4] M. S Andersen, J. Dahl, and L. Vandenberghe. CVXOPT: A Python package for convex optimization, version 1.1. 6. *Available at cvxopt. org*, 2013.

[5] J. Auclair, J. Lemieux, B. Tremblay, and H. Ritchie. Implementation of Newton's method with an analytical Jacobian to solve the 1D sea ice momentum equation. *J. Comput. Phys.*, 340:69–84, 2017.

[6] F. Auger, M. Hilairet, J. M. Guerrero, E. Monmasson, T. Orlowska-Kowalska, and S. Katsura. Industrial applications of the Kalman filter: A review. *IEEE Transactions on Industrial Electronics*, 60(12):5458–5471, 2013.

[7] 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.

[8] A. Berg and L. EB. Eriksson. SAR algorithm for sea ice concentration evaluation for the Baltic Sea. *IEEE Geoscience and Remote Sensing Letters*, 9(5):938–942, 2012.

[9] F. Bouttier and P. Courtier. Data assimilation concepts and methods March 1999. *Meteorological training course lecture series. ECMWF*, 2002.

[10] L. Brucker, D. J. Cavalieri, T.n Markus, and A. Ivanoff. NASA Team 2 sea ice concentration algorithm retrieval uncertainty. *IEEE Transactions on Geoscience and Remote Sensing*, 52(11):7336–7352, 2014.

[11] D. Calvetti, L. Reichel, F. Sgallari, and G. Spaletta. A regularizing Lanczos iteration method for underdetermined linear systems. *Journal of computational and applied mathematics*, 115(1):101–120, 2000.

[12] T. Carrieres, M. Buehner, J-F. Lemieux, and L.T. Pedersen. *Sea Ice Analysis and Forecasting*. Cambridge University Press, 2017.

[13] F. Carsey and B. Holt. Beaufort-Chukchi ice margin data from Seasat: Ice motion. *Journal of Geophysical Research: Oceans*, 92(C7):7163–7172, 1987.

[14] F. D. Carsey. *Microwave remote sensing of sea ice.* American Geophysical Union, 1992.

[15] K. R. Castleman. Digital image processing. *Prentice-Hall, Inc*, 1(996):475–478, 1996.

[16] D. J. Cavalieri, C. L. Parkinson, P. Gloersen, J. C. Comiso, and H. J. Zwally. Deriving long-term time series of sea ice cover from satellite passive-microwave multisensor data sets. *Journal of Geophysical Research: Oceans*, 104(C7):15803–15814, 1999.

[17] A. Caya, M. Buehner, and T. Carrieres. Analysis and forecasting of sea ice conditions with three-dimensional variational data assimilation and a coupled ice-ocean model. *Journal of Atmospheric and Oceanic Technology*, 27(2):353–369, 2010.

[18] J. M. Chambers. *Graphical methods for data analysis.* CRC Press, 1983.

[19] D. A Clausi and H. Deng. Operational segmentation and classification of SAR sea ice imagery. In *IEEE Workshop on Advances in Techniques for Analysis of Remotely Sensed Data, 2003*, pages 268–275. IEEE, 2003.

[20] D. A. Clausi, A. Qin, M. Chowdhury, P. Yu, and P. Maillard. MAGIC: Map-guided ice classification system. *Canadian Journal of Remote Sensing*, 36(sup1):S13–S25, 2010.

[21] J. C. Comiso, D. J. Cavalieri, C. L. Parkinson, and P. Gloersen. Passive microwave algorithms for sea ice concentration: A comparison of two techniques. *Remote sensing of Environment*, 60(3):357–384, 1997.

[22] J. C. Comiso and K. Steffen. Studies of antarctic sea ice concentrations from satellite data and their applications. *Journal of Geophysical Research: Oceans*, 106(C12):31361–31385, 2001.

[23] G. B. Crocker and T. Carrieres. The Canadian Ice Service digital database: History of data and procedures used in the preparation of regional ice charts. *Contract Report No. 00-02, Ballicater Consulting Ltd Ottawa*, 2000.

[24] W. T. Crow and E. F. Wood. The assimilation of remotely sensed soil brightness temperature imagery into a land surface model using ensemble Kalman filtering: A case study based on ESTAR measurements during SGP97. *Advances in Water Resources*, 26(2):137–149, 2003.

[25] R. Daley. *Atmospheric data analysis*. Number 2. Cambridge University Press, 1993.

[26] D. Dee, S. Uppala, A. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. Balmaseda, G. Balsamo, P. Bauer, et al. The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 137(656):553–597, 2011.

[27] D. P. Dee. Bias and data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 131(613):3323–3343, 2005.

[28] J. S. Denker and Y. Lecun. Transforming neural-net output levels to probability distributions. In *Advances in neural information processing systems*, pages 853–859, 1991.

[29] A. Der Kiureghian and O. Ditlevsen. Aleatory or epistemic? Does it matter? *Structural Safety*, 31(2):105–112, 2009.

[30] D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[31] A. M. Ebtehaj and E. Foufoula-Georgiou. On variational downscaling, fusion, and assimilation of hydrometeorological states: A unified framework via regularization. *Water Resources Research*, 49(9):5944–5963, 2013.

[32] A. M. Ebtehaj, E. Foufoula-Georgiou, G. Lerman, and R. L. Bras. Compressive earth observatory: An insight from AIRS/AMSU retrievals. *Geophysical Research Letters*, 42(2):362–369, 2015.

[33] G. Evensen. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans*, 99(C5):10143–10162, 1994.

[34] G. Evensen. *Data assimilation: the ensemble Kalman filter*. Springer Science & Business Media, 2009.

[35] M. Fily and D. A. Rothrock. Sea ice tracking by nested correlations. *IEEE Transactions on Geoscience and Remote Sensing*, 5:570–580, 1987.

[36] G. M. Flato. The thickness variable in sea-ice models. *Atmosphere-Ocean*, 36(1):29–36, 1998.

[37] E. Foufoula-Georgiou, A. M. Ebtehaj, S. Q. Zhang, and A. Hou. Downscaling satellite precipitation with emphasis on extremes: A variational l1-norm regularization in the derivative domain. *Surveys in Geophysics*, 35(3):765–783, 2014.

[38] A. Fowler and P. Jan Van Leeuwen. Observation impact in data assimilation: the effect of non-Gaussian observation error. *Tellus A: Dynamic Meteorology and Oceanography*, 65(1):20035, 2013.

[39] D. Freedman and P. Diaconis. On the histogram as a density estimator: L2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 57(4):453–476, 1981.

[40] M. A. Freitag, N. K. Nichols, and C. J. Budd. Resolution of sharp fronts in the presence of model error in variational data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 139(672):742–757, 2013.

[41] H. Fu, M. K. Ng, M. Nikolova, and J. L. Barlow. Efficient minimization methods of mixed l2-l1 and l1-l1 norms for image restoration. *SIAM Journal on Scientific computing*, 27(6):1881–1902, 2006.

[42] Y. Gal and Z. Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016.

[43] Y. Gal and L. Smith. Sufficient conditions for idealised models to have no adversarial examples: a theoretical and empirical study with Bayesian neural networks. *arXiv preprint arXiv:1806.00667*, 2018.

[44] J. Gascard, K. Riemann-Campe, R. Gerdes, H. Schyberg, R. Randriamampianina, M. Karcher, J. Zhang, and M. Rafizadeh. Future sea ice conditions and weather forecasts in the arctic: Implications for arctic shipping. *Ambio*, 46(3):355–367, 2017.

[45] G. Gaspari and S. E. Cohn. Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society*, 125(554):723–757, 1999.

[46] Z. Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452, 2015.

[47] L. Girard, S. Bouillon, J. Weiss, D. Amitrano, T. Fichefet, and V. Legat. A new modeling framework for sea-ice mechanics based on elasto-brittle rheology. *Ann. Glaciol.*, 52(57):123–132, 2011.

[48] F. Goerlandt, J. Montewka, W. Zhang, and P. Kujala. An analysis of ship escort and convoy operations in ice conditions. *Safety science*, 95:198–209, 2017.

[49] G. H. Golub and U. Von Matt. Tikhonov regularization for large scale problems. In *Workshop on scientific computing*, pages 3–26, 1997.

[50] C. Haas, S. Hendricks, H. Eicken, and A. Herber. Synoptic airborne thickness surveys reveal state of Arctic sea ice cover. *Geophysical Research Letters*, 37(9), 2010.

[51] C. Haas and S. E. Howell. Ice thickness in the Northwest Passage. *Geophysical Research Letters*, 42(18):7673–7680, 2015.

[52] C. Haas, J. Lobach, S. Hendricks, L. Rabenstein, and A. Pfaffling. Helicopter-borne measurements of sea ice thickness, using a small and lightweight, digital EM system. *Journal of Applied Geophysics*, 67(3):234–241, 2009.

[53] D. Hall. *Remote sensing of ice and snow*. Springer Science & Business Media, 2012.

[54] R. Hall and D. Rothrock. Sea ice displacement from Seasat synthetic aperture radar. Technical report, DTIC Document, 1981.

[55] P. C. Hansen. Analysis of discrete ill-posed problems by means of the l-curve. *SIAM review*, 34(4):561–580, 1992.

[56] P. C. Hansen, J. G. Nagy, and D. P. O'Leary. *Deblurring images: matrices, spectra, and filtering*, volume 3. Society for Industrial and Applied Mathematics (SIAM), 2006.

[57] P. C. Hansen and D. P. OLeary. The use of the L-curve in the regularization of discrete ill-posed problems. *SIAM Journal on Scientific Computing*, 14(6):1487–1503, 1993.

[58] D. Haverkamp, L. K. Soh, and C. Tsatsoulis. A dynamic local thresholding technique for sea ice classification. In *Proceedings of IGARSS'93-IEEE International Geoscience and Remote Sensing Symposium*, pages 638–640. IEEE, 1993.

[59] R. Hecht-Nielsen. Theory of the backpropagation neural network. In *Neural networks for perception*, pages 65–93. Elsevier, 1992.

[60] W. D. Hibler. A dynamic thermodynamic sea ice model. *Journal of Physical Oceanography*, 9:815846, 1979.

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

[62] Q. A. Holmes, D. R. Nuesch, and R. A. Shuchman. Textural analysis and real-time classification of sea-ice types using digital SAR data. *IEEE Transactions on Geoscience and Remote Sensing*, GE-22:113–120, 1984.

[63] C. Johnson, N. K. Nichols, and B. J. Hoskins. Very large inverse problems in atmosphere and ocean modelling. *International journal for numerical methods in fluids*, 47(8-9):759–771, 2005.

[64] S. J. Julier and J. K. Uhlmann. New extension of the Kalman filter to nonlinear systems. In *AeroSense'97*, pages 182–193. International Society for Optics and Photonics, 1997.

[65] L. Kaleschke, N. Maaß, C. Haas, S. Hendricks, G. Heygster, and R. Tonboe. A sea-ice thickness retrieval model for 1.4 GHz radiometry and application to airborne measurements over low salinity sea-ice. *The Cryosphere*, 4(4):583–592, 2010.

[66] L. Kaleschke, X. Tian-Kunze, N. Maaß, M. Mäkynen, and M. Drusch. Sea ice thickness retrieval from SMOS brightness temperatures during the Arctic freeze-up period. *Geophysical Research Letters*, 39(5), 2012.

[67] R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering*, 82(1):35–45, 1960.

[68] R. E. Kalman and R. S. Bucy. New results in linear filtering and prediction theory. *Journal of Fluids Engineering*, 83(1):95–108, 1961.

[69] E. Kalnay. *Atmospheric modeling, data assimilation, and predictability*. Cambridge University Press, 2003.

[70] J. Karvonen. C-band sea ice SAR classification based on segmentwise edge features. In *Geoscience and remote sensing new achievements*. InTech, 2010.

[71] J. Karvonen. A comparison of two c-band sar ice/open water algorithms. In *SeaSAR 2010*, volume 679, 2010.

[72] J. Karvonen, M. Similä, 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.

[73] A. Kendall. *Geometry and Uncertainty in Deep Learning for Computer Vision*. PhD thesis, University of Cambridge, 2018.

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

[75] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[76] 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.

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

[78] A. S. Komarov and M. Buehner. Improved retrieval of ice and open water from sequential RADARSAT-2 images. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–9, 2019. doi: 10.1109/TGRS.2018.2886685.

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

[80] C. König Beatty and D. M. Holland. Modeling landfast sea ice by adding tensile strength. *J. Phys. Oceanogr.*, 40:185–198, January 2010.

[81] I. Kubat, M. Sayed, and M. H. Babaei. Analysis of besetting incidents in Frobisher Bay during 2012 shipping season. In *Proceedings of the International Conference on Port and Ocean Engineering Under Arctic Conditions*, 2013.

[82] S. Kullback and R. A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[83] W. Lahoz, B. Khattatov, and R. Menard. Data assimilation and information. In William Lahoz, Boris Khattatov, and Richard Menard, editors, *Data Assimilation: Making Sense of Observations*, pages 3–12. Springer Berlin Heidelberg, 2010.

[84] C. L. Lawson and R. J. Hanson. *Solving least squares problems*. SIAM, 1995.

[85] S. W. Laxon, K. A. Giles, A. L. Ridout, D. J. Wingham, R. Willatt, R. Cullen, R. Kwok, A. Schweiger, J. Zhang, C. Haas, et al. CryoSat-2 estimates of Arctic sea ice thickness and volume. *Geophysical Research Letters*, 40(4):732–737, 2013.

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

[87] 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.

[88] J.F. Lemeiux, B. Tremblay, S. Thomas, Sedlacek J., and L.A. Mysak. Using the preconditioned Generalized Minimum RESidual (GRMES) method to solve the sea-ice momentum equation. *Journal of Geophysical Research*, 113:doi:10.1029/2007JC004680, 2008.

[89] J. Lemieux, C. Beaudoin, F. Dupont, F. Roy, G. C. Smith, A. Shlyaeva, M. Buehner, A. Caya, J. Chen, T. Carrieres, et al. The Regional Ice Prediction System (RIPS): verification of forecast sea ice concentration. *Quarterly Journal of the Royal Meteorological Society*, 142(695):632–643, 2016.

[90] J. Lemieux, D. A. Knoll, M. Losch, and C. Girard. A second-order accurate in time IMplicit-EXplicit (IMEX) integration scheme for sea ice dynamics. *J. Comput. Phys.*, 263:375–392, 2014.

[91] W. H. Lipscomb, E. C. Hunke, W. Maslowski, and J. Jakacki. Ridging, strength, and stability in high-resolution sea ice models. *J. Geophys. Res.*, 112(C03S91), 2007.

[92] H. Liu, H. Guo, and L. Zhang. SVM-based sea ice classification using textural features and concentration from RADARSAT-2 dual-pol ScanSAR data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(4):1601–1613, 2015.

[93] D. Long. Radar, Scatterometers. In Eni G. Njoku, editor, *Encyclopedia of Remote Sensing*, pages 532–535. Springer, 2014.

[94] U. Loptien and L. Axell. Ice and AIS: ship speed data and sea ice forecasts in the Baltic Sea. *The Cryosphere*, 8:2409–2418, 2014.

[95] A. C. Lorenc. Analysis methods for numerical weather prediction. *Royal Meteorological Society, Quarterly Journal*, 112:1177–1194, 1986.

[96] D. J. MacKay. A practical Bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

[97] MANICE. Manual of standard procedures for observing and reporting ice conditions, 2005.

[98] W. N. Meier, G. K. Hovelsrund, B. E. van Oort, J. R. Key, K. M. Kovaks, C. Michel, C. Haas, M. A. Granskog, S. Gerland, D. K. Perovich, A. Makshtas, and J. D. Reist. Arctic sea ice in transformation: A review of recent and observed changes and impacts on biology and human activity. *Review of Geophysics*, 51:185–217, 2014.

[99] W. N. 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.

[100] M. Mohammadi-Aragh, H. F. Goessling, M. Losch, N. Hutter, and T. Jung. Predictability of Arctic sea ice on weather time scales. *Scientific Reports*, 8(6514):doi:10.1038/s41598–017–24660–0, 2018.

[101] R. M. Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

[102] S. Nihashi, K. I. Ohshima, T. Tamura, Y. Fukamachi, and S. Saitoh. Thickness and production of sea ice in the Okhotsk Sea coastal polynyas from AMSR-E. *Journal of Geophysical Research: Oceans*, 114(C10), 2009.

[103] NSIDC: National Snow and Ice Data Center. Submarine upward looking sonar ice draft profile data and statistics, version 1. http://dx.doi.org/10.7265/N54Q7RWK, 1998. updated 2006.

[104] S. Ochilov and D. A. Clausi. Operational SAR sea-ice image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 50(11):4397–4408, 2012.

[105] C. Oliver and S. Quegan. *Understanding synthetic aperture radar images*. SciTech Publishing, 2004.

[106] K. Raney. Radar, Altimeters. In Eni G. Njoku, editor, *Encyclopedia of Remote Sensing*, pages 525–532. Springer, New York, NY, 2014.

[107] A. H. Renner, S. Gerland, C. Haas, G. Spreen, J. F. Beckers, E. Hansen, M. Nicolaus, and H. Goodwin. Evidence of Arctic sea ice thinning from direct observations. *Geophysical Research Letters*, 41(14):5029–5036, 2014.

[108] R. Ressel, A. Frost, and S. Lehner. A neural network-based classification for sea ice types on X-band SAR images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(7):3672–3680, 2015.

[109] L. Robertson and C. Persson. On the application of four dimensional data assimilation of air pollution data using the ajoint technique. In *Air Pollution Modeling and Its Application IX*, pages 365–373. Springer, 1992.

[110] D. A. Rothrock and A. S. Thorndike. Geometric properties of the underside of sea ice. *Journal of Geophysical Research*, 85(C7):3955–3963, 1980.

[111] S. Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[112] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533, 1986.

[113] Y. Sasaki. An objective analysis based on the variational method. *J. Meteor. Soc. Japan*, 36(3):77–88, 1958.

[114] Y. Sasaki. Some basic formalisms in numerical variational analysis. *Monthly Weather Review*, 98(12):875–883, 1970.

[115] K. A. Scott, Z. Ashouri, M. Buehner, L. Pogson, and T. Carrieres. Assimilation of ice and water observations from SAR imagery to improve estimates of sea ice concentration. *Tellus A: Dynamic Meteorology and Oceanography*, 67(1):27218, 2015.

[116] J. A. Screen and I. Simmonds. The central role of diminishing sea ice in recent arctic temperature amplification. *Nature*, 464(7293):1334, 2010.

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

[118] M. Shokr and N. Sinha. *Sea Ice: Physics and Remote Sensing*. John Wiley & Sons, 2015.

[119] L. Soh and C. Tsatsoulis. Unsupervised segmentation of ERS and RADARSAT sea ice images using multiresolution peak detection and aggregated population equalization. *International Journal of Remote Sensing*, 20(15-16):3087–3109, 1999.

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

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

[122] G. Stonebridge, K. A. Scott, and M. Buehner. Impacts on sea ice analyses from the assumption of uncorrelated ice thickness observation errors: Experiments using a 1D toy model. *Tellus A: Dynamic Meteorology and Oceanography*, 70(1):1445379, 2018.

[123] H. A. Sturges. The choice of a class interval. *Journal of the American Atatistical Association*, 21(153):65–66, 1926.

[124] T. Tamura, K. I. Ohshima, T. Markus, D. J. Cavalieri, S. Nihashi, and N. Hirasawa. Estimation of thin ice thickness and detection of fast ice from SSM/I data in the Antarctic ocean. *Journal of Atmospheric and Oceanic Technology*, 24(10):1757–1772, 2007.

[125] M. Thomas, C. Geiger, and C. Kambhamettu. High resolution (400 m) motion characterization of sea ice using ERS-1 SAR imagery. *Cold Regions Science and Technology*, 52(2):207–223, 2008.

[126] A. N. Tikhonov, V. I. Arsenin, and F. John. *Solutions of ill-posed problems*, volume 14. Winston Washington, DC, 1977.

[127] R. T. Tonboe, S. Eastwood, T. Lavergne, A. M. Sørensen, N. Rathmann, G. Dybkjær, L. T. Pedersen, J. L. Høyer, and S. Kern. The EUMETSAT sea ice concentration climate data record. *The Cryosphere*, 10(5):2275–2290, 2016.

[128] G. Wahba. *Spline models for observational data.* SIAM, 1990.

[129] G. Wahba and J. Wendelberger. Some new mathematical methods for variational objective analysis using splines and cross validation. *Monthly weather review*, 108(8):1122–1143, 1980.

[130] L. Wang, K. A. Scott, L. Xu, and D. A. Clausi. Sea ice concentration estimation during melt from dual-pol SAR scenes using deep convolutional neural networks: A case study. *IEEE Transactions on Geoscience and Remote Sensing*, 54(8):4524–4533, 2016.

[131] I. H. Woodhouse. *Introduction to microwave remote sensing.* CRC press, 2017.

[132] Q. Yang, M. Losch, S. N. Losa, T. Jung, L. Nerger, and T. Lavergne. Brief communication: The challenge and benefit of using sea ice concentration satellite data products with uncertainty estimates in summer sea ice data assimilation. *The Cryosphere*, 10(2):761–774, 2016.

[133] Q. Yu and D. A. Clausi. SAR sea-ice image analysis based on iterative region growing using semantics. *IEEE Transactions on Geoscience and Remote Sensing*, 45(12):3919–3931, 2007.

[134] N. Zakhvatkina, A. Korosov, S. Muckenhuber, S. Sandven, and M. Babiker. Operational algorithm for ice–water classification on dual-polarized RADARSAT-2 images. *The Cryosphere*, 11(1):33–46, 2017.

[135] N. Y. Zakhvatkina, V. Y. Alexandrov, O. M. Johannessen, S. Sandven, and I. Y. Frolov. Classification of sea ice types in ENVISAT synthetic aperture radar images. *IEEE Transactions on Geoscience and Remote Sensing*, 51(5):2587–2600, 2013.

[136] H. J. Zwally, J. C. Comiso, C. L. Parkinson, D. J. Cavalieri, and P. Gloersen. Variability of Antarctic sea ice 1979–1998. *Journal of Geophysical Research: Oceans*, 107(C5):9–1, 2002.

[137] M. Zygmuntowska, P. Rampal, N. Ivanova, and L. H. Smedsrud. Uncertainties in arctic sea ice thickness and volume: new estimates and implications for trends. *The Cryosphere*, 8(2):705–720, 2014.

APPENDICES

Appendix A

Generalized objective function

In Section 3.3, the objective functions in equations 3.6 and 3.7 are formulated assuming the background and observation error variances are spatially homogeneous. A more general formulation allowing nonhomogeneous error variances is explained in this Appendix.

We start by considering the l_2 objective function given earlier in equation (3),

$$J(\mathbf{x}) = (\mathbf{x} - \mathbf{x}_b)^T \mathbf{B}^{-1} (\mathbf{x} - \mathbf{x}_b) + (\mathbf{y} - H(\mathbf{x}))^T \mathbf{R}^{-1} (\mathbf{y} - H(\mathbf{x})) \tag{A.1}$$

Following [17] we introduce a change of variable $\xi = \mathbf{B}^{-\frac{1}{2}}(\mathbf{x} - \mathbf{x}_b)$, which means $\mathbf{x} = \mathbf{x}_b + \mathbf{B}^{\frac{1}{2}}\xi$. By substituting $\mathbf{x}$ into J we arrive at the following objective function:

$$J(\xi) = \xi^T \xi + \|\mathbf{R}^{-\frac{1}{2}}(\vec{y} - H(\mathbf{x}_b)) - \mathbf{R}^{-\frac{1}{2}}\mathbf{H}\mathbf{B}^{\frac{1}{2}}\xi\|_2^2 \tag{A.2}$$

By defining $\mathbf{f} = \mathbf{R}^{-\frac{1}{2}}(\mathbf{y} - H(\mathbf{x}_b))$ and $\mathbf{G} = \mathbf{R}^{-\frac{1}{2}}\mathbf{H}\mathbf{B}^{\frac{1}{2}}$ we can write equation (A.2) as

$$J(\xi) = \|\mathbf{f} - \mathbf{G}\xi\|_2^2 + \|\xi\|_2^2, \tag{A.3}$$

For the l_1-l_2-norm method an objective function can be defined that is the same as that for l_2 but with the additional regularisation term,

$$J(\xi) = \|\mathbf{f} - \mathbf{G}\xi\|_2^2 + \|\xi\|_2^2 + \delta\|\mathbf{D}\mathbf{x}_0\|_1. \tag{A.4}$$

This objective function has the same form as that used in the present study, although optimal values of δ would be different since the scaling of the problem is different.

Appendix B

Toy sea ice model

The sea ice model used in the present study is based on the model introduced by [60], and uses a viscous-plastic (VP) rheology. Although some authors have recently proposed new sea ice rheologies (e.g. [47]), almost all current sea ice models are based on the VP rheology. With the VP constitutive law, the ice behaves as a very viscous fluid when the state of stress is inside the yield curve while it deforms plastically when the state of stress reaches the yield curve. The viscous coefficients are capped when the ice is in the viscous regime. The minimum delta value is 2×10^{-9} s^{-1}. This value is so small that the ice is basically modelled as an ideal plastic material.

In 2D, the standard VP elliptical yield curve leads to strong resistance in compression, significant resistance in shear and small resistance in tension. As in [5, 80, 90, 91], the 1D model is obtained by assuming that the v component of the ice velocity and spatial gradients $\partial/\partial y$ are zero. The 1D model is a simplification of the 2D model as one can not simulate failure in shear. However, it still allows one to simulate and study interesting aspects of the VP rheology such as free drift, landfast ice and failures in compression and tension.

The toy sea ice model used in this study solves a momentum equation that can be written:

$$\begin{aligned} &\rho_{ice} a h \frac{\partial u}{\partial t} = \\ &\rho_a a C_a u_a |u_a| + \rho_o a C_o (u_o - u)|u_o - u| + \frac{\partial}{\partial x}\left((\zeta + \eta)\frac{\partial u}{\partial x} - \frac{P}{2}\right), \end{aligned} \tag{B.1}$$

where ρ_{ice} represents the ice density, C_a represents the ice-atmosphere drag coefficient, u_a represents the wind velocity, C_o represents the ice-ocean drag coefficient, u_o represents the ocean current velocity and ρ_a, ρ_o are the air of air and water respectively. The velocity of the water and air were obtained from shallow water models that are coupled to the ice model. The variables in the ice stress term, which is a 1-D version of the stress tensor defined in [60] are

$$\begin{aligned}
\zeta &= \frac{P}{2\Delta} \\
\eta &= \zeta e^{-2} \\
\Delta &= \left|\frac{\partial u}{\partial x}\right| (1 + e^{-2}) \\
P &= P^* h \exp\left(-C(1-a)\right),
\end{aligned} \tag{B.2}$$

where P represents the ice strength, and P^*, C, and e are empirically-derived constants. In addition to the ice momentum equation, there are transport equations for ice thickness and concentration [60]:

$$\begin{aligned}
\frac{\partial a}{\partial t} &= -\frac{\partial ua}{\partial x} + S_a \\
\frac{\partial h}{\partial t} &= -\frac{\partial uh}{\partial x} + S_h
\end{aligned} \tag{B.3}$$

where S_a and S_h are thermodynamic terms governing the ice growth. For the current application both S_a and S_h are zero.

The sea ice equations were discretized using a central differencing scheme with velocity grid points at locations staggered from the thickness, concentration, and viscosity. Atmospheric and oceanic boundary conditions were provided by one-layer shallow water models, tuned to produce qualitatively realistic velocity fields. Periodic boundary conditions were used for the sea ice momentum, concentration and thickness equations as well as for the forcing. Values used for the constants are $P^* = 8000 N/m$, C_a=0.0015, C_w=0.0015, $e = 2$, ρ_a=1.3kg/m^3 and ρ_i=910kg/m^3 and ρ_o=1035kg/m^3. Note that ζ was constrained to be within the range of $\zeta_{min} = 4 \times 10^8 kgs^{-1}$ and $\zeta_{max} = 2.8 \times 10^8 P$ to avoid singularities that can occur for small Δ values and reduce numerical instabilities that can occur for small viscosities [60, 88]. For further details about the ice model see [122].