Predicting Sea Ice Concentration with Calibrated Uncertainty Quantification using Passive Microwave and Reanalysis Data

by

Ray Valencia

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

Waterloo, Ontario, Canada, 2022

Author's Declaration

I hereby declare that I am the sole author of this 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.

Abstract

The adoption of deep learning (DL) techniques in the domain of remote sensing, and specifically sea ice concentration (SIC) prediction, using passive microwave (PM) data and atmospheric climate data has seen a growing interest. Given these predictions, it has been called upon to accompany predictions with their uncertainty, as a means to enhance quality and trustworthiness of results, which can be used in various climate applications in modelling and policy. Though, studies regarding uncertainty quantification (UQ) for SIC prediction has seen little interest. Within DL, there exists a subset of methodologies that work alongside prediction methodologies to effectively quantify uncertainty present within the model, as well as the uncertainty inherent in the data. Among these techniques include Bayesian Neural Networks (BNN's), and heteroscedastic neural networks (HNN's), where the former is used to measure model (epistemic) uncertainty and the latter data (aleatoric) uncertainty. For predicting SIC, and quantifying model and data uncertainty, we propose the use of a combined methodology using a heteroscedastic Bayesian neural network (HBNN) which follows the architecture of a multilayer perceptron (MLP) using PM and atmospheric data. Additionally, we explore the notion of calibration, and related methodologies as a means to evaluate the quality of uncertainties. The advantage of the proposed approach is its data driven nature for prediction and UQ, which is flexible to the context of the given data, such as in space or time. From the results of UQ, it was found that uncertainties vary throughout the seasonal ice cycle, where the months that coincide with melt-onset in the region are susceptible to the highest uncertainties. Additionally, within the study region, uncertainties were scattered, where highest uncertainties were found in areas near or in the marginal ice zone. It was also found that the inclusion of TB's in the feature space are most necessary to produce quality estimates of SIC, and the inclusion of atmospheric variables as input contributed to reduce uncertainty. Finally, when analyzing the effects of calibration on the model, it was found to yield quality and trustworthy predictions of uncertainty.

Acknowledgements

I would like to thank my supervisor, professor Andrea Scott for her continuous help and support over the last two years.

I would also like to thank my thesis committee members, Andrea Scott, Zhao Pan, and Linlin Xu.

An additional thanks as well to members of the Remote Sensing Research Group such as Armina Solemyani, Daniel Sola, and Xinwei Chen for their help, support, and knowledge through all this time.

Last but not least, I would like to thank my family, friends, and partner for providing me the support needed throughout this graduate degree during the heart of COVID-19.

Without all of everyone's support, I would not have been able to succeed through this turbulent stage of life, so thank you again.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Sea ice concentration (SIC) is a variable used by climate scientists to measure the fraction of surface area covered by ice. It is described as a numeric value between zero and one, where zero represents an area devoid of ice, and one, an area fully covered by ice. As a variable, SIC is useful to determine a number of important climate variables such as sea ice surface albedo, ice volume, and ice extent [27], which also act as input for some climate models [22]. Values of SIC are also helpful to nautical navigators on ships that pass through ice covered domains such as in the Arctic and Antarctic, for evaluations in the safety of routes [39].

Passive microwave (PM) sensors onboard satellites are used as a means to monitor sea ice in arctic regions. These sensors measure microwave radiation as emitted from the earth's surface, known as brightness temperature (TB). Using TB data, various SIC algorithms based on empirical tie points can be used to calculate the SIC for a region [38]. PM sensors and SIC algorithms have shown to work well in estimating SIC, but require processing to alleviate weather affects due to overlaps in TB signatures of intermediate SIC and respective weather effects. With recent advances in the realms of machine learning and deep learning, powerful models such as neural networks (NN's) have been used for SIC prediction. They have shown their capability to predict SIC given brightness temperature as input, without the requirements of such processing steps as needed for the previous SIC algorithms [71], [32]. This entices the possibility of such NN models to replace the use of SIC algorithms that traditionally use empirical tie points in future applications.

In the realm of remote sensing, including applications to sea ice, it has been called upon to accompany data records with estimates of their uncertainties, which is necessary to support applications for various policy, climate modelling, and weather prediction ventures [58]. For sea ice remote sensing data, as related to PM-TB, uncertainty quantification has been explored, such as in Tonboe et. al. [75], which assessed uncertainties of SIC estimates as related to tie points and smearing due to sensor footprint, and in Brucker et. al. [10] for uncertainty in NT2 SIC retrieval, but in general has received little attention. Methodologies in neural networks to measure uncertainty have gained traction in recent years [9] [41] [1], and can be easily adapted to suit most architectures. Thus, to estimate SIC, it is sensible to utilize a multilayer perceptron (MLP) neural network architecture with an uncertainty modelling framework.

Additionally, one must take into account the predictions of uncertainty themselves, and evaluate these uncertainties such that they are trustworthy. In deep learning, this notion is called calibration and is a means to measure the quality of uncertainty predictions [28] [45]. Recent work on calibration has been developed in deep learning tasks for classification and regression, primarily in fields such as health sciences and autonomous vehicles. Given the need for uncertainty and respective evaluations of these uncertainties in climate sciences, calibration is also needed.

1.2 Objective

Previous studies on uncertainty quantification for sea ice related applications have shown uncertainty to originate from technical characteristics as related to the measurement or retrieval of data, such as in sensor noise, observational error, or processing errors due to resolution mismatch [10], [75]. This uncertainty is well understood for use in previous methodologies to estimate SIC, but their effect on SIC retrieval with deep learning has not been explored. Thus we propose an approach to not only measure this data uncertainty, but also the uncertainty originating from the model itself, while providing data driven uncertainties, specific to the current data instance which may able to provide insight into local conditions.

The objective of this thesis is then to predict sea ice concentration, quantify epistemic (model) and aleatoric (data) uncertainty, and produce a calibrated model utilizing brightness temperature data obtained from passive microwave sensors, and geophysical atmospheric climate data as input. To estimate sea ice concentration, we propose a deep learning model based on a multilayer perceptron architecture (MLP). As we frame the problem of estimating SIC to be that of solving a non-linear regression problem, the choice of MLP is appropriate. To estimate epistemic and aleatoric uncertainty, we propose the use of the Bayes by backprop method combined with heteroscedastic loss, as these methods translate well to MLP architectures. To calibrate the model, we utilize auxiliary interval predictors, posing a methodology that uses two models to perform both prediction and interval estimation that together produce quality uncertainty estimates needed for a calibrated model.

We also explore the outcome of different sources of training label data to the model, independently testing both training label data from SIC values obtained from bootstrap (BT) and the enhanced nasa team (NT2) algorithms to the model. Additionally, we introduce combinations of geophysical climate variables as input features to the model, and analyze their effects on SIC predictions. Finally, we explore the predictions with respect to uncertainty for the seasonal variations in data.

1.3 Thesis Outline

The remainder of the thesis is outlined as follows. Chapter 2 discusses background information on the topics of sea ice concentration, retrieval methods for sea ice concentration, deep learning, uncertainty, and calibration. Chapter 3 discusses the various data used in the study, as well as the area of interest. The next chapters, (4 and 5), provides technical details of the methodologies used, and subsequent experimental setup. Following this, chapters 6 and 7, describe the experiments and the subsequent analysis of such experiments respectively. To conclude, chapter 8 discusses and summarizes the results, and provides avenues for future work.

Chapter 2

Background

2.1 Sea Ice concentration

Sea ice concentration (SIC) is defined as the measurement of sea ice area relative to the total area in some location [51]. SIC has a value between 0 and 1, where 0 indicates the absence of sea ice, also known as areas of open water and values of 1 indicate full sea ice cover, also known as consolidated ice. The zone of sea ice which transitions from open water to consolidated ice is known as the marginal ice zone (MIZ) and is characterized by SIC values between 0.15 and 0.85. The portion of the MIZ directly bordering that of open water is also known as the ice edge.

2.1.1 Sea Ice Extent

An alternative measure of sea ice is the sea ice extent (SIE). As opposed to SIC, which is a unitless value of sea ice in some area, SIE is a measurement of the approximate area of ocean where sea ice is present. The threshold for what is considered sea ice is typically any area which has SIC greater than or equal to 0.15 or 15%. SIE and SIC are directly related, where SIE can be calculated given SIC data and the spatial resolution of such data.

2.2 Sea Ice Concentration Retrieval

To estimate SIC, several remote sensing methodologies have been used, based off optical imagery, synthetic aperture radar (SAR) imagery and passive microwave (PM) data [76],

[11]. These methodologies use satellite sensors due to their ability to obtain information in vast spatial regions. The advantage of optical imagery lies in the straightforward interpretation of sea ice due to strong contrasts of albedo between sea ice and open water. The downfall of such images though is the obstruction of clouds, which are especially present in the Arctic.

On the other hand, SAR is an active sensor that measures the backscattered signal from the Earth's surface. These measures are generally in the low-frequency portion of the electromagnetic spectrum and are thus not affected by atmospheric moisture nor cloud cover. SAR sensors have high spatial resolution, approximately 50-100 meters (m). SAR imagery from SAR sensors can often be difficult to interpret due to the presence of speckle noise and SAR's sensitivity to both imaging geometry and properties of the surface.

PM sensors, on the contrary, measure microwave radiation emitted from the earth's surface, and at low frequencies, are also not affected by cloud cover. PM sensors do not measure SIC directly, and instead first measure the brightness temperatures (TB) emitted from the earth's surface, then calculate SIC via retrieval algorithms such as the bootstrap (BT) algorithm or enhanced NASA team (NT2) algorithm. PM-TB based SIC estimation data from these algorithms are open source and are readily accessible to the general populous, though may have some pitfalls. SIC estimates are negatively affected by many factors, influencing their accuracy. These include atmospheric weather effects on the sea ice surface and open water, as well as the presence of surface melt [31], [5], [55]. These can then be further complicated due to seasonal/monthly changes of these weather effects [3]. To correct these erroneous effects, the algorithms often use weather filters [26], [13] but have been shown to not only remove weather effects but the ice itself [38]. An alternative approach is to correct the brightness temperatures before using them in the retrieval algorithm [75], [4], [3].

2.3 Deep Learning

Machine learning (ML) research is a continuously growing field in the domain of computer science and has been of interest for the last half-century, attributed to algorithms that can adapt, learn and predict through data. A specific subset of ML, known as deep learning (DL) is specifically of interest as of late, as well as the chosen tool of DL, deep neural networks (DNN's). The interest in NN's can be attributed to the universal approximation theorem, stating that for any function, there exists a NN that can represent and solve said function [36]. Such a concept is exciting, given the numerous problems the scientific community is interested in solving. Here, the concept of deep refers to the model architecture being partially composed of a series of layers where each layer has a set of weights that need to be learned. The minimum architecture that neural networks use consist of an input layer, an output layer, and a variable amount of layers in between the input and output. With recent advances in computational hardware, software, data volume, and data availability, research and production of DNN's has flourished, as they have demonstrated their ability to solve difficult problems and better learn patterns in data as opposed to other ML approaches [50].

2.4 Uncertainty Quantification in Deep Learning

The results obtained from NN's are useful, but are taken without considering whether the results are trustworthy. Such a measure of trustworthiness is the notion of uncertainty. NN's are increasingly used in decision making processes, and such the requirement to provide uncertainty estimates have seen a rises of interest in various domains including health sciences [44], computer vision [41], automated vehicles [20], remote sensing [29], [6] and many more [1]. Additionally, if uncertainty quantification is possible in our models, we can then aim to reduce uncertainty and increase confidence in predictions. In the context of machine learning, uncertainty can be categorized into epistemic and aleatoric uncertainty.

2.4.1 Epistemic Uncertainty

Epistemic uncertainty is described as the uncertainty that stems from a lack of knowledge in a system. If such a systems is a NN model, then the lack of knowledge stems from the inability of the model to predict an output. As this type of uncertainty is attributed to lack of knowledge, it can be improved by producing a model that better encapsulates the problem, or to add more knowledge, in the form of data into the model. Methods to measure epistemic uncertainty include Monte Carlo (MC) Dropout [23], Bayesian NN's [9], and Deep Ensembles [48].

2.4.2 Aleatoric Uncertainty

In contrast, aleatoric uncertainty is the uncertainty that originates from the intrinsic randomness of observations. In a NN, these observations are analogous to the input, i.e. our data of the model, which in this SIC application originate from the instruments that record our data. This cannot be reduced in the same manner as the epistemic uncertainty, but can be reduced by increasing dimensionality of input feature space [37]. The aleatoric uncertainty can be further categorized into heteroscedastic and homoscedastic aleatoric uncertainty [49]. Homoscedastic uncertainty is the uncertainty originating from noise that is assumed to be identical for all points in the data. Heteroscedastic uncertainty on the other hand is the uncertainty when the noise is assumed to be variable across all points in the data.

2.4.3 Model Calibration

Uncertainties are useful when assessing the predictions of the model, but in some uncertainty predicting models, they can fail to capture the true uncertainty in the model. To validate the quality of such uncertainty measurements, model calibration is used. Studies [45], [48], [52] have shown models that capture uncertainty, such as Bayesian NN's are inherently non-calibrated, and fail to capture the true distributions of data. In the case of uncertainty quantification for regression, a model is said to be calibrated if the observed confidence level matches exactly with an expected confidence level. Here the observed confidence level is the observed ground truth values that fall within a predictive interval (PI) as produced by the prediction of the model at a specific expected confidence level. For example, for a PI produced at an expected confidence level of 95%, we should expect to find 95% of ground truth values are contained within this PI. If at this same PI (produced at an expected confidence level of 95%) there is found to only have 80% of ground truth values contained within it, then the model is uncalibrated.

2.5 Related Work

Deep learning methodologies for predicting SIC has gained popularity in the last decade, an example of which is the use of deep convolutional neural networks (CNN's) that have demonstrated their ability to produce significant improvements to SIC estimates from SAR data during both melt and freeze-up periods as compared to passive microwave data [77], [78]. These studies used ice charts as training labels, but similar approaches have been done using PM data as training labels, which have shown success [19], [66]. Among other approaches in deep learning methodologies for SIC endeavors, some studies have utilized historical SIC data to forecast monthly estimates of SIC and use architectures such as multilayer perceptron's (MLP) [14], LSTM (Long short term memory) [14], and deep ensembles [42]. Encompassing studies concerning estimating SIC with deep learning methodologies using PM data, some have used additional features with TB, showing improvements in accuracy while alleviating short comings produced by PM based algorithms and their respective algorithms [71], [32]. These additional features include various atmospheric geophysical variables such as windspeed and air temperature, typically obtained via a reanalysis data set, such as from the European Reanalysis Agency-5 (ERA-5) dataset [34].

To assess the reliability and trustworthiness from the results obtained from these deep learning methodologies, one can perform uncertainty quantification (UQ) [1], [25]. Uncertainty quantification has been explored for SIC products, such as for NT2 SIC [10], but for endeavors concerning UQ using deep learning methodologies for sea and lake ice remote sensing, they have only been recently explored. An example of which is in Asadi et. al. [6], where they proposed a methodology utilizing MLP's to quantify epistemic and aleatoric uncertainty in classification and detection of ice and water in SAR imagery. In the study, introducing uncertainty helped in reducing misclassification of ice and water in the domain. Additionally, quantification of aleatoric uncertainty from a convolutional neural network was incorporated for lake ice mapping using SAR images in Saberi et. al [70], which found the addition of incorporating uncertainty helped to improve water and ice mapping.

On its own, UQ can help to address trustworthiness of ML predictions, but lack in evaluations to validate the quality of the uncertainties themselves. In recent years, studies have used the notion of model calibration as a means to evaluate predictions of models, as well as predictions of uncertainty. Methods to calibrate models have been explored in both the context of classification [28], [33], [61] and regression tasks [45], [48], [74]. Most of the methodologies have been applied to problems such as medical diagnosis [53], [67], [73] and autonomous vehicles [64]. Recent applications in remote sensing have been explored for producing well calibrated uncertainties in precipitation type classification [63], and calibrated uncertainty quantification for estimating canopy height [2]. In sea ice related applications, a calibration approach based on temperature scaling was used in Anderssen et. al [5] for their probabilistic deep learning ensemble method for sea ice forecasting.

Chapter 3

Dataset and Study Area

3.1 Study Area and Timeframe

The area of interest for this study falls in the North Eastern portion of Canada, a region covered primarily in seasonal ice, which is ice that completely melts in the summer and forms again in the winter. The area was chosen given that it is a region experiencing declines in SIC [47], with increases in shipping activity [65], and contains part of the Tallurutiup Imanga National Marine conservation area, which is an important habitat for marine mammals and seabirds [47, 30]. The area comprises the whole of Baffin Bay, Davis Strait, some of the Labrador Sea, and most of Nares Strait towards the the Lincoln Sea (Figure 3.1). Baffin Bay is located between Greenland in the east, Baffin Island in the West, and the Davis strait directly south. To encapsulate relatively recent trends within this region, we explore the use of the year 2020 and 2021. Specifically, we train the models based on features (as described later) on the year of 2020, and perform predictions (inference) for the year of 2021.

3.2 Brightness Temperature Data

For this study, we utilize two independent sets of brightness temperature data as input for experiments. The first are from the Advanced Microwave Scanning Radiometer 2 (AMSR2) Brightness temperatures (TB). This TB data is derived from the Japanese Aerospace Exploration Agency (JAXA) AMSR2 dataset. The data is comprised of swatch observations from six frequencies. The six frequencies are: 6.9, 10.7, 18.7, 23.8, 36.5, and 89 Ghz, at

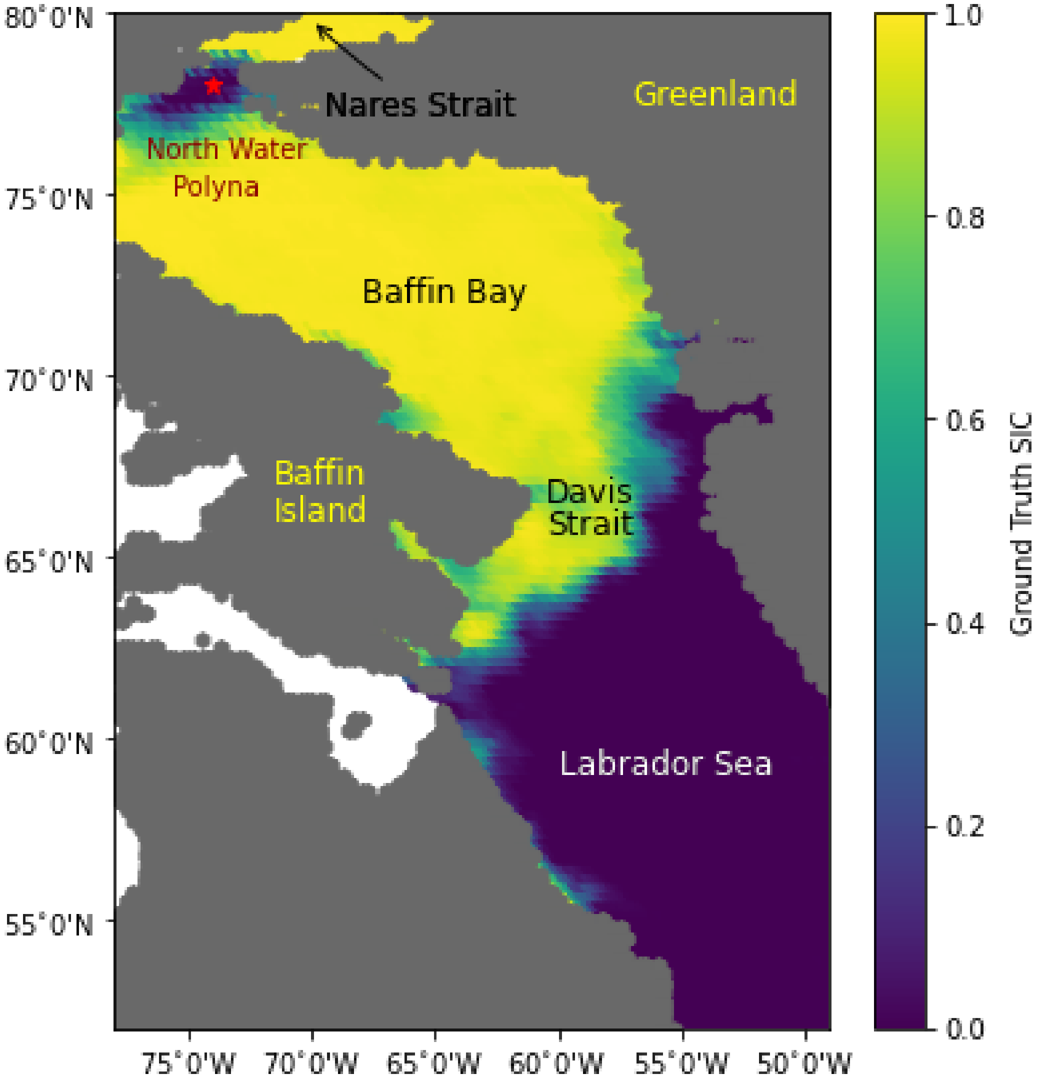

Figure 3.1: Study area map for May of 2021, which shows the average SIC as calculated by the NT2 algorithm, having a nominal gridded resolution of 12.5 km

both horizontal and vertical polarization. The spatial resolution for each channel differs, with instantaneous field of views (IFOV) of 35 x 62 km, 24 x 42 km, 14 x 22 km, 11 x 19 km, 7 x 12 km, and 3 x 5 km respectively [56]. As the original data are swath observations, processing is done by the NSIDC to map observations onto a 12.5 km polar stereographic grid. This is done by using a method which takes the sum and average of data samples that fall within the same grid cell (also known as a drop-in-the-bucket method). Data from AMSR2 are available from July 2nd, 2012 to present.

The second set of brightness temperature data are the special sensor microwave imager/sounder (SSMIS) TB data which are from the SSMIS sensor onboard the Defense Meteorological Satellite Program (DMSP) F17 platform. Brightness temperatures are measured at four frequencies, and are the 19.3, 22.2, 37.0, and 91.7 Ghz channels, with horizontal and vertical polarization's for each channel available. Here the IFOV for each channel are 42 x 70 km for both 19 and 22 Ghz channels, 28 x 44 km for the 37.0 channel, and 13 x 14 km for the 91 Ghz channel [57]. As the raw data are satellite swath observations, processing is done by the NSIDC to map the observations onto a 25 km polar stereographic grid using a similar drop-in-the-bucket method as above. The data from F17 SSMIS sensor is available from June 12th, 2006 to present.

3.3 Sea Ice Concentration Data

For the present study, we use two sets of sea ice concentration (SIC) data from two sources.

3.3.1 Bootstrap Algorithm

The bootstrap (BT) algorithm uses brightness temperature (TB) observations from 37 horizontal (H), 37 vertical (V), and 19.3V Ghz (which for ease we will refer to hereafter as 19V) channels from the F17 SSMIS sensor to estimate SIC [17], [15]. Scatterplots between the two channels are created, where two non-linear clusters are identified. Most data points in consolidated ice regions where ice concentration is greater than 95% are clustered in a common area, where a line along this area is inferred from a regression analysis, and is known as the line A-D. Additionally, most points that correspond to open water are clustered in a different area of the scatter plot, and can be represented as the line known as O-W. The points closest to O usually correspond to lowest brightness temperatures in the plot. Take for example an arbitrary ice surface represented by the point I along A-D. Different concentrations of this ice type are represented by data points along the line O-I.

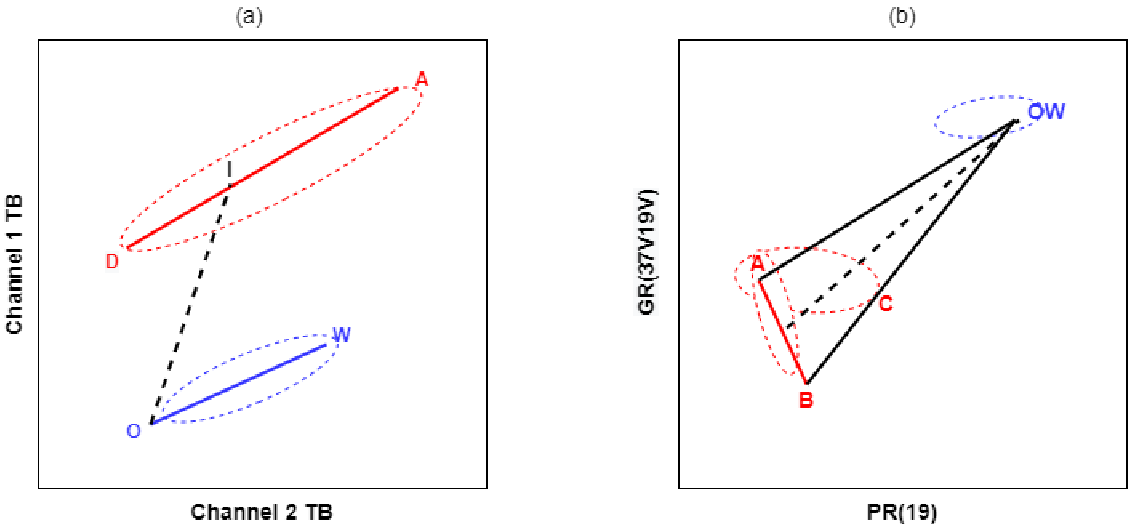

Figure 3.2: Schematic diagram of the technique used in two SIC algorithms. Here red ellipses simulate group of scatter points corresponding to consolidated ice, while blue ellipse simulate groups of scatter points corresponding to open water (a) Schematic for the BT algorithm. Points in the consolidated ice region where SIC $> 95\%$, are represnted as the (red) line A-D. Most of the ice free and/or open water points are clustered along the (blue) line O-W. The (black) dotted line I-O measures SIC relative to the distance to A-D or O-W. (b) Schematic for the NT algorithm. The tie points A, B, and OW, correspond to first year ice, multi year ice, and open water respectively. The (red) A-B line corresponds to 100% SIC. The distance from the point OW to the line A-D, represented as the (black) dashed line, is a measure of the SIC. Laslty, the group of points clustered at C are points with significant surface effects.

Given some data point along the line I-O, the distance from such point to I (or O) is a measure of the sea ice concentration, where SIC = 0 corresponds to points near O, and SIC = 1 corresponds to points closer to or on the line A-D. A schematic diagram of this method is provided in Figure 3.2(a).

Additionally the bootstrap algorithm uses two more plots to derive sea ice concentrations for different regions of ice. Higher TB measurements are typically observed within the ice pack, as opposed to near the ice edge, which helps to delineate the two regions. Then, for TBs within the ice pack, plots of 37H vs 37V (polarization mode) are used to calculate SIC. To calculate SIC near the ice edge boundary, TBs of 37V vs 19V (frequency mode) are plotted, since the combination of 37V and 19V is more sensitive to the ice-water boundary. Additionally, a weather filter is applied to help identify areas where weather related erroneous effects may affect SIC estimates [18].

The SIC from the bootstrap algorithm is available as the Bootstrap Sea Concentration dataset at the NSIDC [15], and has a nomimal gridded resolution of 25 km.

3.3.2 Nasa Team (NT) and Enhanced Nasa Team (NT2) Algorithms

SIC obtained from the the NASA Team (NT) algorithm uses two equations [54]. The first is the polarization ratios of brightness temperatures,

$$PR(v) = \frac{TB(vV) - TB(vH)}{TB(vV) + TB(vH)}, \tag{3.1}$$

and the second is the spectral gradient ratio of brightness temperatures,

$$GR(v1pv2p) = \frac{TB(v1p) - TB(v2p)}{TB(v1p) + TB(v2p)}. \tag{3.2}$$

Here, TB is the brightness temperature at a frequency v, for a polarized component p, i.e. vertical (V) or horizontal (H), where the brightness temperatures are obtained from the AMSR2 sensor [56].

First, the NT algorithm calculates the polarization ratios of the 18.7 GHz brightness temperatures, and plots it against the spectral gradient ratio calculated between the 36.5 GHz vertical (V) and 18.7 GHz vertical (V) brightness temperatures, an example can be seen in Figure 3.2(b). For ease, we denote the 18.7 and 36.5 GHz channels as 19 and 37 Ghz respectively from this point onward. Here, tie points for first-year (A), multi-year ice

(B), and open water (OW) are identified and shown, then a line connecting A-B to OW is subsequently plotted. To measure SIC, the relative distance a point contained on the line connecting A-B to OW, is measured relative to either A-B or OW. The closer the point is to OW, the lower the SIC, while the closer the point is to the line A-B, the higher the the SIC. The primary source of error in the NT algorithm can be attributed to ice surface effects such as glazing and layering [16], which can affect 19 GHz TBH's, underestimating SIC. These significant surface effects are clustered as a group of points, denoted C, away from the 100% ice concentration line A-B. The difference between GR(89V19V) and GR(89H19H), known as ΔGR help to distinguish between pixels of low ice concentration and pixels with significant surface effects.

The NT2 algorithm utilizes a similar basis to that of the NT, but employs a greater complexity than that of the NT to account for surface effects, atmospheric types, and weather related effects, where a short summary of the process is detailed as follows. First, the response of theoretical TB's to different weather conditions are calculated using an atmospheric radiative transfer model [46]. As input, the model uses emissivities of first year ice under winter conditions, as well as various atmospheric profiles with different cloud properties, atmospheric temperatures, and humidity profiles for summer and winter conditions. Following this, theoretical brightness temperatures for all possible ice concentration and weather combinations are calculated, and for each of these solutions the ratios between PR(19), PR(89), and ΔGR are calculated. This creates a "prism" in which each element contains a vector of these three ratios. Next, similar PR(19), PR(89) and ΔGR ratios are calculated from the observed AMSR2 brightness temperatures. The weather corrected ice concentration ratios is found by minimizing the observed ratios to the theoretical (modelled) ratios. The weather corrected ratios are then used to calculate the final SIC via plots of GR(37V19V) and PR(19).

Although the NT2 algorithm produces weather corrected sea ice concentrations, erroneous SIC estimates are still possible. Thus, the NT2 algorithm requires additional methodologies to correct SIC estimates. The first is an atmospheric correction scheme, providing weather corrected SIC through a forward atmospheric radiative transfer (RT) model. This helps to eliminate remaining severe specious ice concentrations in open water, while also applying atmospheric corrections to currently covered ice. For the most severe weather effects, which are present in open ocean, the use of weather filters [26], [13] are needed. In some cases, spurious SIC are found in low latitude locations, so a sea surface temperature (SST) filter from the National Oceanic and Atmospheric Administration (NOAA) are used [12]. Then, to correct for erroneous SIC measures found along coast lines, a land mask is overlayed and applied.

The SIC data using the NT2 algorithm is available at the AMSR2 Unified Daily Brightness Temperatures and Sea Ice Concentration's from the NSIDC [56], which has a nominal gridded resolution of 12.5 km and uses TB's from the from the AMSR2 sensor.

3.4 Reanalysis Data

In this study, 5th generation reanalysis (ERA5) data produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) [34] is utilized for input into the MLP models. The ERA5 dataset is a large collection of various atmospheric, land, and oceanic climate variables, provided from the year of 1979 to present. The data covers most regions of the earth, with a nominal gridded resolution of 30 km. From the ERA5 dataset, we utilize 4 climate variables.

The first climate variables is an aggregation of two original components of the 10-meter meridional-component and 10-m zonal-component of wind. The ERA5 dataset collects the in-situ data from the Drifting Buoy dataset from the world meteorological information system (WMO WIS). The 10-meter meridional-component is the horizontal speed of air moving towards the north at a height of ten metres above the earths surface. On the contrary, the 10-meter zonal-component of wind is the horizontal speed of air moving towards the east. The two components are then transformed into a single measure of the horizontal 10-meter windspeed (WS), by calculating the magnitude between the meridional-component and zonal-component. The second climate variable used is the 2 meter air temperature (AT), which is the temperature of air 2 meters above the surface of land or sea measured in kelvin (K). ERA5 collects and aggregates temperature data from the WMO WIS. The third climate variable is the total column vertically integrated water vapour (WV), which is the total amount of water vapour in a column extending from the surface of the earth to the top of the atmosphere measured in millimeters (mm). The fourth and last climate variable is the total column cloud liquid water (LW), which is the amount of liquid water contained within cloud droplets in a column extending from the surface of the earth to the top of the atmosphere and is measured in mm. Both VW and LW are collected by ERA5 from satellite sensor data.

3.5 Data organization and processing

For this study, the atmospheric variables and brightness temperatures are used as input features into the neural network model. We use data in the study area for the full years of 2020 and 2021, where 2020 is used in training of the model, and 2021 is used in inference.

SIC data is used solely as labels during model training for the 2020 dataset, while for the 2021 dataset, they are utilized purely as a means for visual comparison between the ground truth and predicted SIC. For the 2020 set, data is split such that 80% is used for training, and 20% is used for validation.

As brightness temperatures are used in the algorithms for estimating SIC we incorporate brightness temperatures as an input feature. For models which use BT SIC values as training labels, the TB obtained from the SSMIS sensor (which are the TBs used in calculating said SIC in the original BT algorithm) are used as the input features in the model. Conversely, models which use the NT2 sic values as training labels utilize the TB from AMSR2 as input features into the model, as this TB was the same TB used in the original NT2 algorithm for SIC estimation. As mentioned previously, both AMSR2 and SSMIS have TB's measured at various channels. Although higher frequency channels such as 89 GHz (from AMSR2) and 91.7 GHz (from SSMIS) have the finest spatial resolutions, they are sensitive to atmospheric water vapor and cloud liquid water [72]. Thus, we choose the 37 GHz frequency from AMSR2 and 37 GHz frequency from SSMIS at both horizontal and vertical polarization's due to its lower sensitivity [59] and its availability across historical passive microwave sensors, enabling the method to be more easily extended to climate data records.

Additionally, as discussed in Chapter 2.5, the use of climate variables are common when applying corrections to sea ice concentration, specifically WS, WV, and LW, which have been used in atmospheric physical models. Furthermore, AT has shown to have strong correlation to sea ice concentrations, and thus is a variable of interest that may help produce better results. Note that data from the ERA5 dataset has an hourly temporal resolution. To be consistent with the daily gridded TB's from AMSR2 and SSMIS, we take the average of the hourly ERA5 data over each day in 2020 and 2021. An example of the input features consisting of TBs from AMSR2 and the ERA5 atmospheric climate variables can be found in Figure 3.3 for the month of May in 2021.

Finally, as the data from AMSR2, SSMIS, and ERA5 have different spatial resolutions, a nearest neighbour interpolation scheme is applied to upsample the lower resolution 30 km ERA5 data to either the 12.5 km AMSR2 TB and NT2 SIC values or 25 km SSMIS TB and BT SIC values, depending on which model is utilized in experiments.

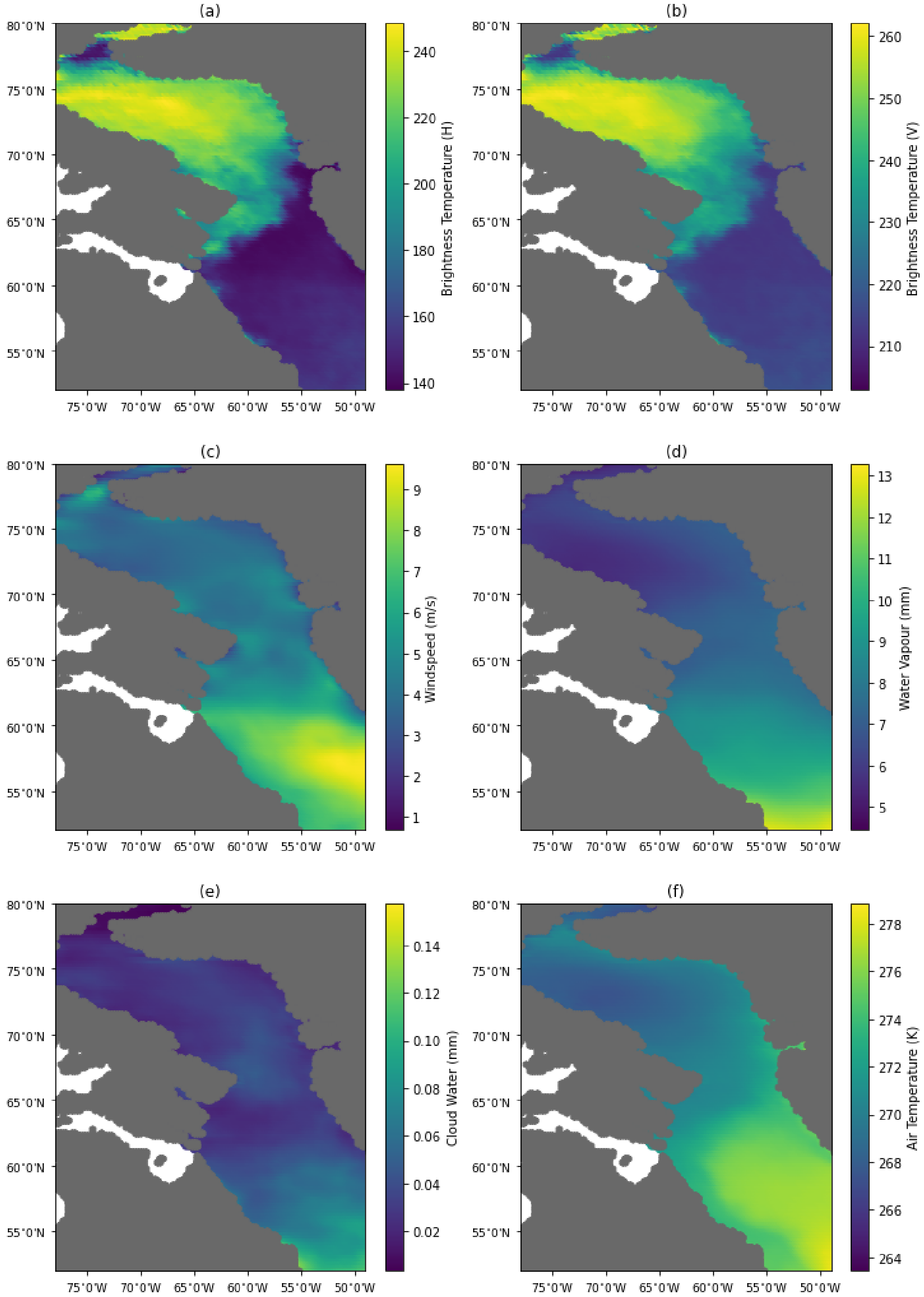

Figure 3.3: Maps of input features used in the testing set for NT2 based models averaged over the month of May 2021. The features are (a) Brightness Temperature (H) (TBH), (b) Brightness Temperature (V) (TBV), (c) Windspeed (WS), (d) Water Vapour (WV), (e) Cloud Water (CW), (f) Air Temperature (AT). TB's shown are obtained from the AMSR2 sensor.

Chapter 4

Methodology

The methodology of this paper is comprised of parts to predict sea ice concentration, measure epistemic and aleatoric uncertainty, and provide a calibrated model. For choice of model architecture, we choose a multilayer-perceptron. To measure the epistemic uncertainty in our model, we utilize the method known as Bayes By Backprop (BBB) [9] to produce a Bayesian neural network (BNN). Following this, aleatoric uncertainty is captured by transforming a BNN into a heteroscedastic BNN (HBNN) which uses the heteroscedastic loss function [41]. Finally the method for calibration is based off the method of Auxiliary Interval Predictors by [74]. We point to the original papers for full technical details but provide a brief overview of each methodology here.

4.1 Multilayer Perceptron

The problem of predicting sea ice concentrations (SIC) can be thought of as a non linear regression problem. Multilayer perceptron (MLP) neural network models are a popular choice for solving non linear regression problems [69], [36] and thus may be well suited to predict SIC. MLP models are a type of feedforward neural network characterized by several layers of neurons, which at minimum consist of an input layer, output layer, and a variable number of hidden layers in between the input and output (Figure 4.1). Given ($\mathbf{x}$, $\mathbf{y}$), where $\mathbf{x}$ is the samples of input data consisting of n features, i.e. $\mathbf{x} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$, and $\mathbf{y}$ are the labels used for training, then the number of neurons in the input layer correspond to the n features of $\mathbf{x}$. The inputs are forward propagated through the layers of the network until they reach the output layer. Note that the number of hidden layers, number of hidden layer neurons, and the neurons in the output layer vary and are defined by the user, usually dependent on the problem at hand. For this instance, assuming that the only goal is to predict sea ice concentration, the number of output neurons for the MLP is 1. Note, as discussed later in Chapter 4.3, this output may be altered to produce not just predictions of SIC, but uncertainty.

Within each layer of the model, the neurons from each forward propagation of the input are calculated as,

$$\nu_i^{l+1} = \phi(\sum_{i=1}^{m} w_i^l \nu_i^l + b^l). \tag{4.1}$$

Here, the model is comprised of m neurons at some hidden layer l. The variable ν_i^{l+1} represents the neuron at the current layer of forward propagation $l+1$, and ν_i^l is the connecting neuron to ν_i^{l+1} in the previous hidden layer l. The w_i^l term corresponds to the weight of connections between neurons, while b^l is the bias within the layer, and ϕ is a nonlinear activation function.

When the input reaches the output layer through forward propagation and an output $\hat{\mathbf{y}}$ is produced, the loss of the network can be calculated between $\hat{\mathbf{y}}$ and the training labels $\mathbf{y}$, using a loss function. For regression problems, a common loss function utilized is the mean squared error (MSE) loss, which is defined as,

$$L_{MSE} = \frac{1}{N} \sum_{i=1}^{N} ||y_i - \hat{y}_i||^2, \tag{4.2}$$

where y_i and $\hat{y}_i$ represent the N elements of the vectors $\mathbf{y}$ and $\hat{\mathbf{y}}$ respectively. Finally, this loss is backpropagated through the network [69], updating the weights between neurons in the network. This process of alternating forward and back propagation are repeated to optimize the network in an attempt to get the best possible output, represented by the lowest loss, or until a stopping criteria is invoked.

4.2 Bayes by Backprop

The weights and biases of a NN, such as an MLP, are provided as single values [69], [79], [36], [35]. One way to quantify the uncertainty of the NN models is to define probabilistic distributions on the weights and biases of the model, such that each weight is defined by a variance and mean (assmuming the form of a Gaussian distribution). An example of

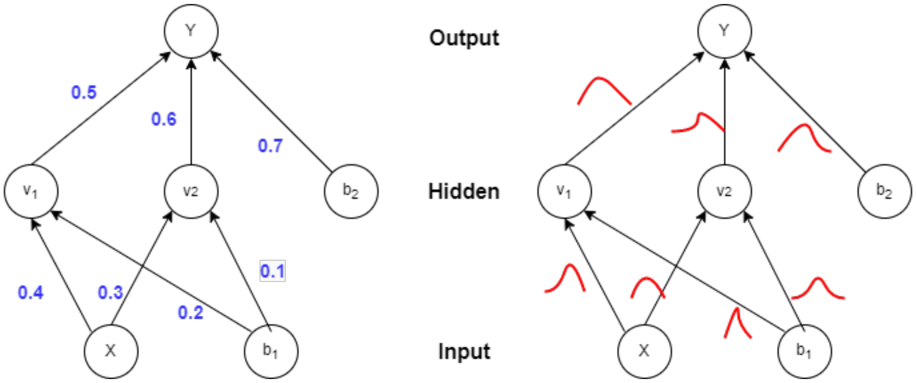

Figure 4.1: Left: Neural network with point values (blue) as weights. Right: (Bayesian) Neural network with distributions (red) representing the weights.

the differences between point and distribution defined weights can be seen in Figure 4.1. Bayes by backprop suggests a method to achieve this, using a variational approximation to Bayesian inference.

Bayes Theorem and Inference: We first introduce Bayes theorem [40], which is defined as,

$$P(\mathbf{w}|D) = \frac{P(D|\mathbf{w}), P(\mathbf{w})}{P(D)}, \tag{4.3}$$

where $P(\mathbf{w}|D)$ is the posterior, $P(\mathbf{w})$ the prior, $P(D|\mathbf{w})$ the likelihood, and $P(D)$ is the scaling factor. Here, $\mathbf{w}$ is the vector of weights in our neural network, and D is the data, characterized by sets of vectorized $(\mathbf{x}, \mathbf{y})$ input/output pairs. For the present problem, it is of interest to find $P(\mathbf{w}|D)$, which is the probability for some weights to be calculated after the data has been seen, and gives the maximum a posteriori (MAP) point estimates of our weights $\mathbf{w}$. This can be further approximated as the following,

$$P(\mathbf{w}|D) \approx P(D|\mathbf{w})P(\mathbf{w}),$$

as the denominator of Equation 4.3, i.e. the scaling factor, is not a function of $\mathbf{w}$, and hence does not change the posterior with respect to $\mathbf{w}$.

If we instead had a distribution over weights, as opposed to point estimates, as in the MAP estimation, we could make predictions that take weight uncertainty into account. To produce such distributions upon our weights $\mathbf{w}$, we apply Bayesian inference, which looks to calculate the posterior predictive distribution (PPD), for some arbitrary new output $\mathbf{y}^*$, and input $\mathbf{x}^*$,

$$P(\mathbf{y}^*|\mathbf{x}^*, D) = \int P(\mathbf{y}^*|\mathbf{x}^*, \mathbf{w})P(\mathbf{w}|D)d\mathbf{w}.$$

Unfortunately, the calculation of the PPD is intractable, stemming from the intractability of the posterior. And thus, the functional form of neural networks does not allow for exact integration of Bayesian inference.

Variational Inference: BBB proposes a variational approximation of the posterior as opposed to the exact Bayesian approach. This variational approximation uses a simpler distribution as a proxy for the true posterior distribution where the proxy posterior distribution (or variational distribution), parameterized by θ, is as close to the true posterior distribution as possible. To find the optimal theta (θ^*) that minimizes the difference between the two, the use of the Kullback-Leibler (KL) divergence is employed between the variational distribution, and the true posterior $P(\mathbf{w}|D)$. Calculating θ^* can be formalized as,

$$\theta^* = arg \min_{\theta} KL[q(\mathbf{w}|\theta)||P(\mathbf{w}|D)], \tag{4.4}$$

where the KL divergence is a distance measure between two probability distributions. When the KL divergence is minimized, the variational distribution is close to or equal to the true distribution. Equation 4.4 can be further broken down as the following loss function for use in a neural network,

$$F(D, \theta) = KL[q(\mathbf{w}|\theta)||P(\mathbf{w})] - E_{q(\mathbf{w}|\theta}[logP(D|\mathbf{w})], \tag{4.5}$$

where the loss function of Equation 4.5 is comprised of two parts, a prior dependent part (known as the complexity cost) minus a data dependent part (also known as the likelihood cost). Rearranging the components of the KL divergence term using a Monte Carlo approximation [9] allows for a new form to be rewritten as,

$$\begin{aligned} F(D, \theta) = E_{q(\mathbf{w}|\theta)}&[\log q(\mathbf{w}|\theta)] \\ &- E_{q(\mathbf{w}|\theta)}[\log p(\mathbf{w})] \\ &- E_{q(\mathbf{w}|\theta)}[\log P(\mathbf{w})]. \end{aligned}$$

Notice that all three terms are expectations with respect to the variational posterior. An unbiased Monte Carlo sampling scheme [69], drawing N_{mc} samples directly from the variational posterior can be used to approximate the expectation in each term. Doing this, the new form of this loss function, and the form that is utilized in the BBB methodology is,

$$L_{BBB} = \frac{1}{N_{mc}} \sum_{j=1}^{N_{mc}} [\log q(\mathbf{w}_j|\theta) - \log P(\mathbf{w}_j) - \log P(D|\mathbf{w}_j)]. \tag{4.6}$$

This form of the loss function is derived and used due to its simplicity of understanding of each component of log posterior, log prior, and log likelihood, as well as a simpler implementation in practice. Here $\mathbf{w}_j$ denotes the jth Monte Carlo weight vector drawn from the variational posterior, and N_{mc} the total draws taken. The variance between draws as taken from the variational posterior is the epistemic uncertainty of the model.

Posterior: The variational posterior ($q(\mathbf{w}|\theta)$) is defined to be a Gaussian distribution. For a single weight of w_i, it is shifted by a mean μ_i and scaled by a standard deviation $\sigma_{w(i)}$. Note that the standard deviation vector is parameterised by the softplus function (pointwise) as,

$$\boldsymbol{\sigma_w} = \log{(1 + exp(\boldsymbol{\rho}))}, \tag{4.7}$$

such that $\boldsymbol{\sigma_w}$ is always non-negative. Thus the variational posterior has the parameters $\theta = (\boldsymbol{\mu}, \boldsymbol{\rho})$. Neural networks require the use of a forward pass and backward pass in training to effectively learn and solve a problem. The forward pass draws a sample or samples from the variational posterior, a stochastic process, while the backwards (backpropagation) takes the gradients of $\boldsymbol{\mu}$ and $\boldsymbol{\rho}$ and updates said values via an optimizer. Since the forward pass is stochastic in nature, the reparameterization trick [62] is used to effectively backpropagate the two variables. The trick is to first sample a variable $\boldsymbol{\epsilon}$ from a parameter free distribution, and then transform $\boldsymbol{\epsilon}$ with a deterministic function $t(\theta, \boldsymbol{\epsilon})$ for which a gradient can then be properly defined. The algorithm for optimisation of the variational posterior parameters in BBB is described in Algorithm 1. The gradients for the mean and standard deviation can

Algorithm 1 Reparameterization Algorithm

```
Sample $\boldsymbol{\epsilon} \approx N(0, I)$
Let $\mathbf{w} = \boldsymbol{\mu} + \log{(1 + exp(\boldsymbol{\rho}) \circ \boldsymbol{\epsilon}}$
Let $\theta = (\boldsymbol{\mu}, \boldsymbol{\rho})$
Calculate loss using $L_{BBB}$
Calculate the gradients of $\boldsymbol{\mu}$ and $\boldsymbol{\rho}$ as $\Delta\boldsymbol{\mu}$ and $\Delta\boldsymbol{\rho}$
Update the variational parameters
```

be learned by the common backpropagation algorithm in neural networks, and are scaled and shifted appropriately as described above.

Prior: Next, the prior is chosen. It may be intuitive to use a simple Gaussian for the prior, which the original authors of BBB noted, but it was found that choosing a scale mixture of Gaussian densities as the prior proved to produce better results. It is defined as follows,

$$P(\mathbf{w}_j) = \prod_k \pi N(\mathbf{w}_k|0, \sigma_{p1}^2) + (1-\pi)N(\mathbf{w}_k|0, \sigma_{p2}^2), \tag{4.8}$$

where $\mathbf{w}_k$ is the kth weight of the current jth Monte Carlo weight vector. Additionally, each density is defined to have zero mean, with differing variances of σ_{p1}^2 and σ_{p2}^2, and scaled by π. The first mixture component is given a larger variance then the second ($\sigma_{p1}^2 > \sigma_{p2}^2$), while also requiring the second mixture component to have a variance much less than one ($\sigma_{p2}^2 << 1$). This achieves two things: 1) It provides a heavier tailed Gaussian, and 2) it causes many weights to tightly center around zero. This prior is shared among all the weights, making it tractable to use during the optimisation step.

Likelihood: Finally, for the likelihood component, the choice of form is dependent on the problem statement, as it is the component of the loss function most related to the data. For instances of binary classification for example, it may be appropriate to choose the cross entropy loss function. In the case of the non-linear regression problem presented in this paper, mean square error (MSE) loss is an appropriate choice, similarly seen in Equation 4.2,

$$P(D|\mathbf{w}) = P(\mathbf{x}|\mathbf{y}, \mathbf{w}) = L_{MSE}. \tag{4.9}$$

Combining the three components of Gaussian variational posterior, Gaussian mixture prior, and MSE loss (as our likelihood) into the function of Equation 4.6 produces our true loss function.

4.3 Heteroscedastic Loss

In BNN's such as of our iteration of BBB, the data dependent portion of the loss function, i.e. the likelihood cost, is measured by the MSE loss. This output corresponds to a Gaussian distribution assuming a constant (homoscedastic) variance, and is not usually accounted for in the loss function. As this variance is constant, there is no notion of the uncertainty which originates from the data. A proposed method by Kendall et. al. [41] considers non-constant variances as to measure the data uncertainty, or aleatoric uncertainty. To measure the aleatoric uncertainty in a BNN, specifically BBB, one can replace the current MSE loss in the likelihood portion, to that of the heteroscedastic loss,

$$L_{HNN} = P(D|\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\Big(\frac{||y_i - \hat{y}_i||^2}{\hat{\sigma}^2_{a(i)}} + \log(\hat{\sigma}^2_{a(i)})\Big), \tag{4.10}$$

where the heteroscedastic loss is the log of a Gaussian assuming a non-constant variance. Here y_i and $\hat{y}_i$ represent the N elements of the vectors $\mathbf{y}$ and $\hat{\mathbf{y}}$ respectively. The term $\hat{\boldsymbol{\sigma}}_{\boldsymbol{a}}$ is the aleatoric standard deviation vector for the model. This term is obtained from a network which has two output neurons, one vector for the predicted mean $\hat{\mathbf{y}}$, and the other vector for $\hat{\boldsymbol{\sigma}}_{\boldsymbol{a}}$. This vector of $\hat{\boldsymbol{\sigma}}_{\boldsymbol{a}}$ differs from $\boldsymbol{\sigma}_w$, since the latter is the standard deviation which parameterizes the posterior and subsequently all weights $\mathbf{w}$ in the model. In practice, Kendall et. al. suggests to train the network on the log variance,

$$s_i = \log(\hat{\sigma}^2_{a(i)})$$

such that the new proposed loss function, which replaces Equation 4.10 as the likelihood component in BBB as,

$$P(D|w) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}(\exp(-s_i)(y_i - \hat{y}_i)^2 + s_i). \tag{4.11}$$

This is advised, since the training on the log variance is numerically stable when regressing the aleatoric variance, as the loss avoids division by zero. The exponential mapping also allows the regressor to regress on unconstrained scalar values, producing positive values for the variance. Note, NN's that produce this output are sometimes referred to as heteroscedastic NN's (HNN), and as we are transforming the BNN from Bayes by backprop to include this additional architecture, it is appropriate to distinguish it as a heteroscedastic Bayesian NN (HBNN) or heteroscedastic Bayes by backprop (HBBB).

4.4 Auxiliary Interval Predictors

A newer approach for calibration in regression settings was proposed by Thiagarajan et. al. (2019) [74] using what is known as Auxiliary Interval Predictors. The approach utilizes two separate models, a main model and auxiliary (aux) model, where the main model is used for mean prediction of SIC and the aux model for predictive interval estimation. The strategy behind the approach allows for each model to take into account one another's results in an alternative training fashion, as to match the mean predictions to that of the prediction intervals.

First we define the main model (or mean estimaotr) to be $F(\mathbf{x};\Theta)$, and the auxiliary model (predictive interval (PI) estimator) to be $\tau(\mathbf{x};\Phi)$, parameterized by Θ and Φ respectively. Then, the objective function which is to be minimized is,

$$\begin{aligned} &\min_{\Theta} L_F(\Theta;\mathbf{x},\mathbf{y},\tau(\mathbf{x};\Phi^*), s.t., \\ \Phi^* = &\, arg\min_{\Theta} L_\tau(\Phi;\mathbf{x},F(\mathbf{x};\Theta)). \end{aligned} \tag{4.12}$$

In the above equation, the main model $F(\mathbf{x};\Theta)$ produces mean estimates that takes the PI's into account from the auxiliary model $\tau(\mathbf{x};\Phi^*)$. The auxiliary model then takes into account the parameters from the main model for PI estimation, as well as performing the calibration process. The authors [74] note that the task of producing prediction intervals is used to regularize the main task of fitting a predictive model, i.e. producing mean estimates, by constraining its uncertainty estimates to match the estimated intervals. This process as denoted by the authors is referred to as uncertainty matching.

Interval Estimator: The interval estimator operates such that the model works to achieve a calibration level α, which implements the task of calibration as a differentiable function. The loss of the interval estimator L_τ is produced on the current condition of the parameters defined in the mean estimating term, i.e. Θ, and is represented in three portions. The first portion of loss optimizes calibration, measuring the empirical probability of the true target lying in the estimated intervals,

$$L_{emce} = \left|\alpha - \frac{1}{N}\sum_{i=1}^{N}\mathbf{I}[(\hat{y}_i - \delta_i^l) \leq y_i \leq (\hat{y}_i + \delta_i^u)]\right|. \tag{4.13}$$

Here $\hat{y}_i$ is the result of evaluating the mean estimator at x_i, i.e. $F(x_i;\Theta)$. $\mathbf{I}$ is an indicator function implemented as,

$$SIGMOID[\eta(y_i - \hat{y}_i^l)(\hat{y}_i^u - y_i)],$$

where η is some large scaling factor ($\eta \geq 1e6$) and $\hat{y}_i^l$ and $\hat{y}_i^u$ are the lower and upper bounds of output $\hat{y}_i$,

$$\hat{y}_i^l = \hat{y}_i - \delta_i^l$$
$$\hat{y}_i^u = \hat{y}_i + \delta_i^u.$$

The next portion looks to perform width regularization, which is done by matching the widths v_i to the residuals r_i from F,

$$L_{noise} = \sum_{i=1}^{N} \Big|0.5 * v_i - |r_i|\Big|, \tag{4.14}$$

where,

$$v_i = \delta_i^u + \delta_i^l$$
$$r_i = y_i - \hat{y}_i.$$

The third and final portion ensures the widths of the estimated prediction are as tight as possible around the mean prediction, and as formulated as,

$$L_{sharp} = \sum_{i=1}^{N} \Big(|\hat{y}_i^u - y_i| + |y_i - \hat{y}_i^l|\Big). \tag{4.15}$$

Finally, the combined loss function for the interval estimator L_τ is:

$$L_\tau = L_{emce} + \beta_n L_{noise} + \beta_s L_{sharp}, \tag{4.16}$$

where β_n and β_s are penalty terms to scale the effects of their respective loss portion.

Mean Estimator: The mean estimator $F(x;\Theta)$ predicts mean estimates of a target, i.e. the SIC in the context of our study, along with uncertainty estimates. Here the uncertainties are guided by the parameters Φ of the interval estimator τ to ensure high quality estimates of both the mean, and uncertainties. The mean estimator is evaluated by its loss function L_F, which is defined as,

$$L_F = \frac{1}{N} \sum_{i=1}^{N} \Big(\frac{1}{2}\Big(\frac{||y_i - \hat{y}_i||^2}{\hat{\sigma}_{a(i)}^2} + \log\big(\hat{\sigma}_{a(i)}^2\big)\Big) + \lambda_m |\hat{\sigma}_{a(i)} - \gamma \frac{v_i}{2}|\Big), \tag{4.17}$$

where, $\hat{y}_i$ is the result of evaluating the mean estimator at x_i, i.e. $F(x_i;\Theta)$. The mean estimator $F(x;\Theta)$ is designed to be a heteroscedastic neural network, where it returns a mean and standard deviation estimate, as defined above in section 4.3. Notice that Equation 4.17 is similar in form to the heteroscedastic Loss function of Equation 4.10, but with the added penalty scaling portion, which utilizes a penalty term λ_m and a scaling term γ. The scaling term γ is defined to be $\gamma = 1/z_{(1-\alpha_v)/2}$, with α_v being the calibration level achieved for that current instance by L_τ, and z indicates the z-score.

Predictive Interval: For the methodology of Auxiliary Interval predictors, the variances estimated by the mean estimator loss function are used to estimate intervals for calibration. Recall that calibration for regression is based on the notion that the observed confidence level matches with an expected confidence level. The observed confidence level is the proportion of observed ground truth values that fall within the PI, while the expected confidence level is some threshold. Take the toy example presented in Figure 4.2. Here PI's are produced from predictions given some trained model, at an expected confidence level of 0.90 or 90%. We should observe that approximately 90% of ground truth SIC's fall within these PI's for the model to be calibrated. But, only 80% of ground truth SIC's are observed to fall within their PI's at this expected confidence level of 90%, thus the model is said to be non-calibrated. For a model to be "perfectly" calibrated, its observed and expected confidence levels should be equal for all discrete expected confidence levels, $\varsigma_{exp} : [0, 1]$. We now define how the predictive interval is calculated given a single mean prediction of SIC $\hat{y}_i$ as,

$$\begin{aligned} PI_{(i),\varsigma_{exp}} &= [PI^{lower}_{(i),\varsigma_{exp}}, PI^{upper}_{(i),\varsigma_{exp}}] \\ &= [\hat{y}_i - z_{(1-\varsigma_{exp})/2}\frac{\hat{\sigma}_{a(i)}}{M}, \hat{y}_i + z_{(1-\varsigma_{exp})/2}\frac{\hat{\sigma}_{a(i)}}{M}], \end{aligned} \tag{4.18}$$

where z indicates the z-score (a function of the expected confidence level), and M the number of samples taken from the model, which in our case is equal to the number of N_{MC} samples. Then, for these discrete expected confidence levels, the corresponding observed confidence level ς_{obs} is calculated. We use the following formulation, which is to count the empirical frequency that the ground truth SIC y, is contained within the PI of its corresponding predicted SIC $\hat{y}$ for all points i,

$$\varsigma_{obs} = \frac{|\{y_i | PI^{lower}_{(i),\varsigma_{exp}} \leq y_i \leq PI^{upper}_{(i),\varsigma_{exp}}), i = 1, ..., K\}|}{K}, \tag{4.19}$$

where K is the total amount of points. Finally, as a measure to evaluate calibration, a plot of observed vs. expected confidence levels, known as calibration plots [45], [74] (Figure 7.4) are used.

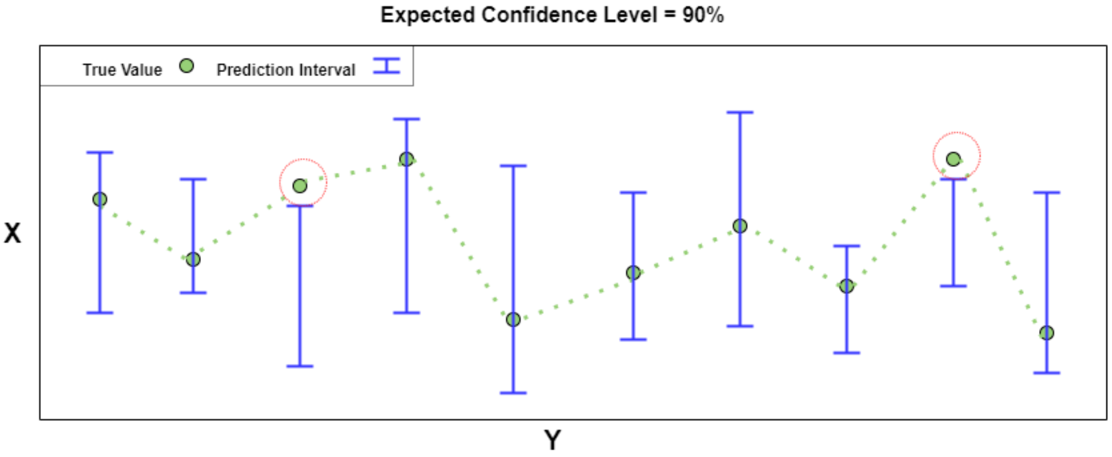

Figure 4.2: Calibrated regression toy problem. Here green points correspond to true (label) values. Blue lines indicates prediction intervals (PIs) as produced by predictions (not shown) from a theoretically trained model at an expected confidence level of 90%. Observe two points (circled in red) not contained within their respective PI's, resulting in an 80% observed confidence level. As the expected confidence level of 90% does not equal the observed confidence level of 80%, the model is said to be non-calibrated for this confidence level.

Calibration Algorithm: The overall algorithm incited for training is described in Algorithm 2. It repetitively performs the mean estimation and interval prediction but in an alternating fashion. The mean estimator takes into account the intervals from the interval prediction for its update, while the interval estimator takes into account the mean prediction. This process is repeated until convergence or until a set amount of the outer epochs has been reached.

Algorithm 2 Calibration using Auxiliary Interval Predictors

```
Input: Labeled Data $\{(x_i, y_i)\}_{i=1}^{N}$,
        Target Calibration Level $\alpha$,
        Epochs $n_{outer}$, $n_{main}$, $n_{aux}$
Initialize: Randomly Initialize $\Theta^*$, $\Phi^*$
for $n_{outer}$ epochs do
    for $n_{main}$ epochs do
        Compute intervals $\delta_i^u, \delta_i^l = \tau(x_i; \Phi^*)$
        Compute Loss Function $L_F$ (Equation 4.17)
        Update $\Theta^* = arg\min \Theta L_F$
    end for
    for $n_{aux}$ epochs do
        Get predictions $\hat{y}_i$ from $F(x_i; \Theta^*)$
        Compute Loss Function $L_\tau$ (Equation 4.16)
        Update $\Phi^* = arg\min \Phi L_\tau$
    end for
end for
Output: Trained Mean and Interval Estimators $F$ and $\tau$
```

Chapter 5

Experimental Setup

We culminate the parts described in the methodology which allows for a well calibrated model that produces accurate mean estimates and measures both epistemic and heteroscedastic aleatoric uncertainty.

By utilizing Bayes by Backprop, we can transform non-Bayesian models such as an MLP into Bayesian models (i.e. BNN) such that the weights are represented as probabilistic distributions, but the original MLP architecture stays the same. For our method, we follow [71] and define an MLP to have 4 hidden layers between the input and output layers, with each hidden layer comprised of 10 neurons. Between each layer of the model, a ReLU activation function is used and for choice of optimiser, we utilise Adam. Through trial and error, we found that a learning rate of 0.001 worked best combined with mini batches comprised of 1000 samples per batch. The data is split into train/validation/test sets as described in the data section. Note that this MLP architecture is chosen for its efficiency and quality predictions of SIC [71]. A summary of the MLP values can be found in Table 5.1.

For BBB, hyperparameter tuning was also required. For N_{mc} it was found that values between 10-50 provided robust results, with any values greater than 50 only providing minimal improvements, but greater computational requirements. Thus an appropriate value of N_{mc} to set to best balance computational efficiency and garner good results was 30. Prior parameters (σ_{p1}, σ_{p2}, π) were chosen to be 1.0, 0.01 and 0.5 respectively, based on Blundell et. al. [9]. Subsequently, the mean $\boldsymbol{\mu}$ of the variational posterior is initialized to be a vector of 0's for all weight values, such that the distribution is centered, while $\boldsymbol{\rho}$ is set to -2 [9]. A summary of the BBB hyperparameters can be found in Table 5.2. To then measure aleatoric uncertainty, we replace the MSE loss of BBB, with the heteroscedastic loss, producing a HBNN, or HBBB specifically.

In the case of auxiliary interval predictors for calibration, it requires the use of HNN's, which fits well with our HBBB iteration. We thus define two models, a main HBBB and auxiliary HBBB where both models retain the same Bayesian architecture. When training is completed, the main HBBB model is now said to be calibrated, and as such is denoted as CHBBB, where C stands for calibrated.We also experiment with the hyperparameters of this method. For choice of calibration level α, we train such that a value of 0.95 is achieved as seen in Equation 4.13. Theoretically, this indicates that the model is fully calibrated for all observed confidence levels at and below 0.95. Next, both β terms are set to 0.1. This is similar to the values set by the original authors of the paper [74]. The β terms help to balance the contributions of each component of the loss function, such that the largest effect on learning is the L_{emce} term (Equation 4.13). From experiments utilizing higher β values (greater than 0.5), it was found that uncertainty estimates for aleatoric uncertainty were greatly constrained which would negatively affect mean predictions, leading to higher RMSE's. Thus relaxing these terms to values less than or equal to 0.1 allowed for better learning of the mean and aleatoric uncertainty. We found that it was also important to scale λ_m appropriately. If the value is set too high (i.e. $\lambda_m \geq 1$) the aleatoric uncertainty can be constrained to a very narrow range, which is a similar effect as the scaling of the β penalty terms. Thus for λ_m, a value of 0.1 is set. The incorrect learning stems from the L_F loss function (Equation 4.17). Observe from (Equation 4.17), the last term of the loss function which includes the v_i term. If the width of v_i is too narrow and becomes a single constant value for all values i, the aleatoric uncertainty $(\hat{\sigma}_a)$ can inflate in learning, and eventually converge to aleatoric uncertainty values that are constant for all values and represent no variability. As both loss functions in Equations 4.16 and 4.17 require parameters learned and transferred from one another, it can produce a cycle of incorrect learning, and thus these adjustments for both the β terms and λ term are necessary. A summary of the calibration methodology hyperparameters are found in Table 5.3.

Lastly, we perform experiments as analyzed in the results and analysis section, with models using varying combinations of input features and training labels. For ease, we provide a table as illustrated in Table 5.4. Here MLP denotes models trained only on an MLP, which can predict SIC but are not capable to measure epistemic nor aleatoric uncertainty. Models denoted as BBB are Bayes by backprop based, which predicts SIC and measures epistemic uncertainty. Models with the added H, as in $HBBB$ utilize the heteroscedastic loss to produce a heteroscedastic neural network, predicting SIC and measuring both epistemic and aleatoric uncertainty, while models with C, such as $CHBBB$ are models calibrated with the auxiliary interval predictors method. Subscripts that follow the main model name, indicate the training label used, where $NT2$ indicates models using

NT2 SIC values as training labels, and *BT* indicates models using BT SIC values as training labels.

Table 5.1: Hyperparameter Summary for MLP architecture

Input Dimensions	6
Hidden Layers	4
Neurons in each layer	10
Output Dimension	2
Learning Rate	0.001
Batch Size	1000
Activation Function	ReLU
Optimizer	Adam

Table 5.2: Hyperparameter Summary for BNN

Initial Posterior Mean (μ)	0.0
Initial Posterior Rho (ρ)	-2
Prior sigma 1 (σ_{p1})	1.0 [9]
Prior sigma 2 (σ_{p2})	0.01 [9]
Prior pi (π)	0.5 [9]
Monte Carlo Samples (N_{mc})	30

Table 5.3: Hyperparameter Summary for Auxiliary Interval Predictors

Lambda match penalty term (λ_m)	0.1
Beta noise penalty term (β_n)	0.1
Beta sharp penalty term (β_s)	0.1
Calibration level (α)	0.95 [74]
Scaling factor (η)	1e6
Outer epochs (n_{outer})	10
Main model epochs (n_{main})	5
Auxiliary model epochs (n_{aux})	5

Table 5.4: Models and Data Summary

Model Name	**Uncertainty Measure**	**SIC Label**	**TB sensor**	**Input Features**
MLP_{NT2}	N/A	NT2	AMSR2	TBH, TBV, AT, WS, WV, LW
BBB_{NT2}	Epistemic	NT2	AMSR2	TBH, TBV, AT, WS, WV, LW
$HBBB_{NT2}$	Epistemic, Aleatoric	NT2	AMSR2	TBH, TBV, AT, WS, WV, LW
$CHBBB_{NT2}$	Epistemic, Aleatoric	NT2	AMSR2	TBH, TBV, AT, WS, WV, LW
$HBBB_{BT}$	Epistemic, Aleatoric	BT	SSMIS	TBH, TBV, AT, WS, WV, LW
$CHBBB_{BT}$	Epistemic, Aleatoric	BT	SSMIS	TBH, TBV, AT, WS, WV, LW
$CHBBB_{NT2-only-tb}$	Epistemic, Aleatoric	NT2	AMSR2	TBH, TBV
$CHBBB_{NT2-no-tb}$	Epistemic, Aleatoric	NT2	N/A	AT, WS, WV, LW

Chapter 6

Experiment Descriptions

6.1 Experiments on Monthly Data

We look to perform experiments as they pertain to monthly observations over an annual cycle. This is done to explore monthly variations of RMSE, epistemic uncertainty, and aleatoric uncertainty as related to periods within the study region such as in times of freeze up, melt, and summer months where little ice is present. For this analysis, we utilize inference results from the 2021 testing dataset on the fully trained and calibrated heteroscedastic neural network of Bayes by backprop, utilizing SIC training labels calculated via the NT2 algorithm, which we denote as $CHBBB_{NT2}$ and the model utilizing training labels calculated via the BT algorithm denoted as $CHBBB_{BT}$.

6.2 Comparison Between Methodologies

A few methodologies are chosen for comparison to the proposed methodology of $CHBBB_{NT2}$ on the basis of RMSE, epistemic uncertainty, and aleatoric uncertainty. They are the base model MLP (MLP_{NT2}), Bayes by backprop (BBB_{NT2}), and a heteroscedastic NN using BBB ($HBBB_{NT2}$). This is done to measure changes as greater complexity is introduced to each model. Here the models are trained on the full year of 2020 data, and inference is performed for the full year of 2021. As per the results of the previous set of experiments, May was found to have the results with the greatest RMSE and total uncertainty overall. It would be of interest to show the capabilities of each model based on the least performing month.

6.3 Spatial Experiments using a Calibrated Model with NT2 SIC as training labels

As predictions of SIC and values of epistemic and aleatoric uncertainty are tied to unique latitude and longitude values, it is possible to plot these values to evaluate their spatial distribution. It is also of interest to explore areas of the study region showing the most accurate estimates of SIC, as well as areas most (and least) susceptible to higher epistemic and aleatoric uncertainties. Given numerous models, we first focus on the $CHBBB_{NT2}$ model as a baseline experiment, but perform experiments and analysis on other models in subsequent sections.

6.4 Experiments on Non-Calibrated and Calibrated Models

Next, spatial behaviour of calibrated models and how they differ from their non-calibrated counterparts are then examined. We perform spatial evaluation similar to those performed for $CHBBB_{NT2}$ and compare it to the non-calibrated $HBBB_{NT2}$. Additionally, to measure calibration, a calibration plot is constructed where points of expected confidence level vs. the observed confidence level are plotted. Here, the measure of calibration using the approach of auxiliary interval predictors ($CHBBB_{NT2}$) is compared to that of the non calibrated $HBBB_{NT2}$. This same comparison is performed between the uncalibrated $HBBB_{BT}$ and $CHBBB_{BT}$ models.

6.5 Spatial Experiments on NT2 vs. BT as SIC Training Labels

Subsequently, the effects that the use of different SIC training labels have on the prediction of SIC as well as quantification of uncertainty are considered. As described in Chapter 3, we utilize two sets of data as input with varying training labels. The first set of data uses brightness temperatures obtained from the AMSR2 sensor combined with 4 atmospheric variables. This set uses SIC training labels calculated via the NT2 algorithm. The second dataset uses brightness temperatures obtained from the SSMIS sensor combined with the same 4 atmopsheric variables as input to the model. Here, the SIC training labels are calculated from the BT algorithm. Similar to the $NT2$ case, we train two models, the $CHBBB_{BT}$ and its non-calibrated counterpart $HBBB_{BT}$.

6.6 Feature Experiments

6.6.1 Experiments on input feature combinations

We explore the spatial effects of features as they pertain to SIC estimates, RMSE, epistemic uncertainty and aleatoric uncertainty. We look at two groupings of input features, one where only TB's are used as input, and the other where no TB's, i.e. only atmospheric variables, are used. As SIC retrieval algorithms utilize TB's we look at what specific contributions these TB's make to SIC predictions within the proposed model, independent of the additional features. Conversely, the same can be said for the contributions of the atmospheric variables, and whether they are necessary, or if simply using only TB's are sufficient. Lastly, the annual cycle of the model using only TB's as input is examined in comparison to the annual cycle of the models previously using TB's combined with atmospheric variables as input.

6.6.2 Experiments on specific features

Subsequently specific details of the features as they are related to the monthly trends is analyzed. We look at the monthly trends of each feature as related to the annual cycle, with respect to the cumulative total (epistemic plus aleatoric) uncertainty. Here the cumulative uncertainty is the summation of the total uncertainty over the whole region for a given month. In addition, we look at the correlation between input features and SIC training labels.

Chapter 7

Results and Analysis

7.1 Monthly Observations

First, we observe the RMSE's for the $CHBBB_{NT2}$ model as presented Fig 7.1 (top). A pattern is recognized that shows the gradual increasing of RMSE starting in January, which then peaks in May, sharply declines in the summer and September, culminating in a slow increase again in October, November, and December. The lowest RMSE values correspond to the months of October and September, while the highest were found to be the months of April, May, and June. We observe a similar pattern to that of the $CHBBB_{BT}$ model, but with a larger RMSE magnitude in all months, with significantly higher values in the months of January to May, and December.

When observing the uncertainty of both the $CHBBB_{NT2}$ model and $CHBBB_{BT}$, a similar pattern can be seen for the total uncertainty (where the total uncertainty is equal to the sum of the epistemic and aleatoric uncertainties) as compared to the RMSE, but with smaller difference in magnitude. Taking only the epistemic uncertainty into account, it is apparent that the highest epistemic uncertainty (model uncertainty) is in the late spring and summer months of June, and July for both models, but with slightly higher epistemic uncertainties for the $CHBBB_{BT}$ model. The aleatoric uncertainty (data uncertainty) on the other hand, peaks in May for both models, but is similar to the epistemic uncertainty where the $CHBBB_{BT}$ model has slightly higher aleatoric uncertainty consistently throughout the year. We note that the highest values of both RMSE and total uncertainty occur in May for the $CHBBB_{NT2}$ model, and for the $CHBBB_{BT}$ model May is indicated to have the highest uncertainty and second highest RMSE. This coincides with the melt onset in this region [8]. However, this does not coincide with the trend seen in SIE, where SIE sharply decreases in the month of April. This could be attributed to the advection of sea ice out of the domain [7]. The lowest values of RMSE and total uncertainty for both models coincide with times of low SIC, even in the absence of weather filters and/or TB corrections as used in previous studies [26], [13], [75], [4], [3].

Next, we observe that when comparing periods of freeze up (i.e. November and December) to winter conditions (i.e. January, February, and March), the former has comparably lower RMSE than the latter, consistent with both models, albeit different magnitudes. From examination of the ice charts during Nov and Dec this is the time of freeze-up. At this time the initial ice cover is thin in the northern portion of Baffin Bay, and then becomes thicker as the ice cover also expands to cover more of the region, which indicates that the model works well for thin ice periods. Lastly, the aleatoric uncertainty is similar among the months in the same seasonal time frame (i.e. freeze up periods vs times of melt).

7.2 Evaluation of Methods

Evaluating the methods used to esimate SIC (Table 7.1), we can observe a gradual decrease in the RMSE with increasing model complexity, showing an improvement in the accuracy of the estimated SIC relative to the SIC for the test year. The RMSE is is highest for the base model MLP_{NT2} returning an average RMSE over the whole domain for May 2021 of 0.283. To the contrary, the lowest RMSE is for the calibrated heteroscedastic neural network using Bayes by Backprop, $CHBBB_{NT2}$, with a value of 0.235, nearly a 20% improvement in RMSE. Next, the epistemic uncertainty is nearly 0.066 for the BBB or 6.6%, while $HBBB_{NT2}$ and $CHBBB_{NT2}$ reduce this uncertainty to 1.5% and 1.8% respectively. Finally, the aleatoric uncertainty between models only slightly varies.

7.3 Spatial Analysis of Calibrated Model which uses NT2 SIC as training labels

From Figure 7.2(b), the spatial distribution of predicted SIC from the $CHBBB_{NT2}$ model for May 2021, aligns well with the expected spatial distribution of the NT2 SIC training labels (Figure 3.1) for areas of open water in the far south of the domain, as well areas of consolidated ice in the northern portion near Baffin Bay. It also captures the North Water Polynya [21], [60], and the portion of Nares Strait near 80°N. Further, the RMSE

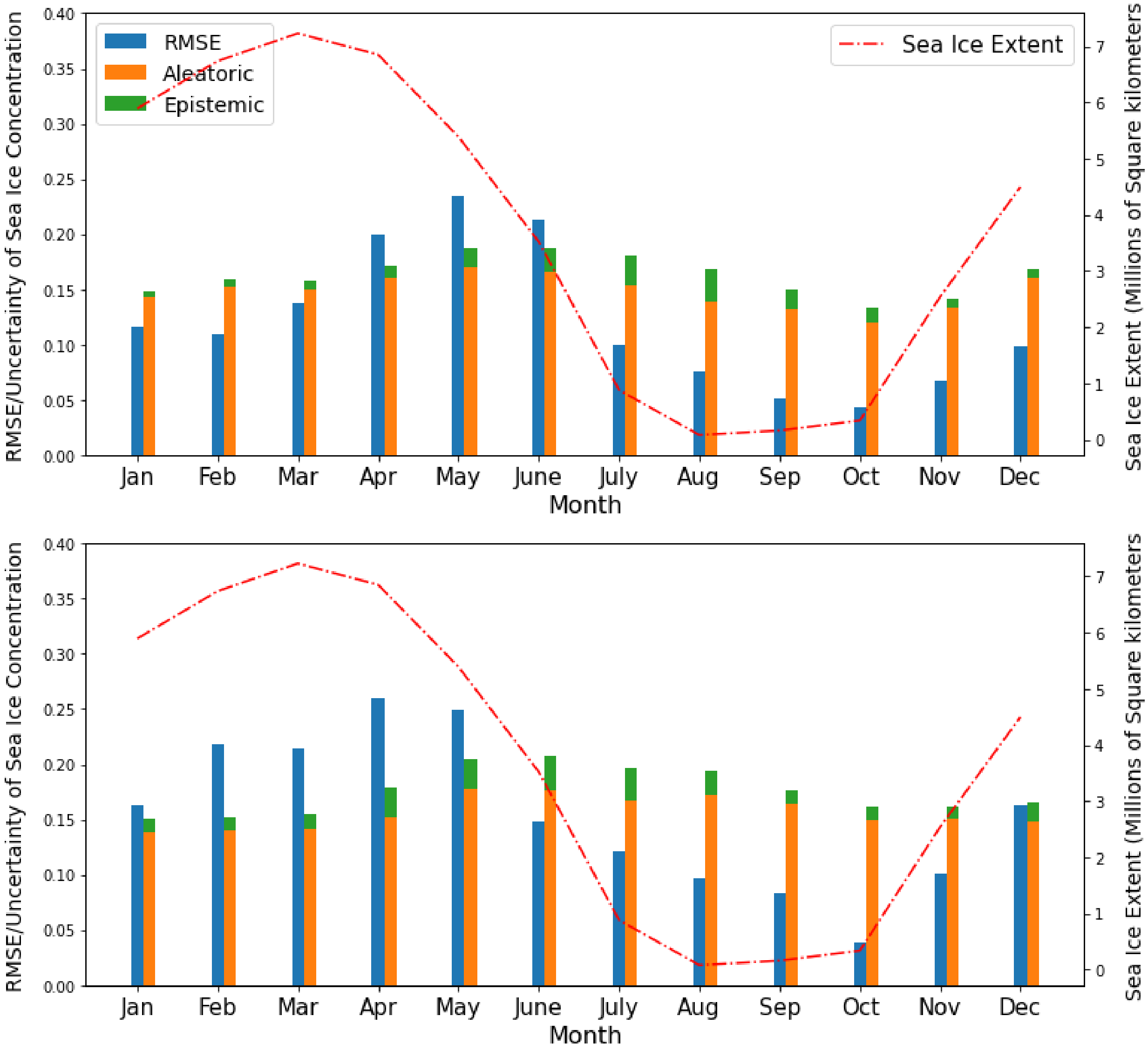

Figure 7.1: Root mean squared error (RMSE), epistemic uncertainty, aleatoric uncertainty using the $CHBBB_{NT2}$ model (top) and $CHBBB_{BT}$ (bottom). The aleatoric and epistemic uncertainty are shown as a stacked bar, representing total uncertainty. The sea ice extent (SIE) values averaged per month for the testing of year of 2021 are also shown for reference to the seasonal cycle. Note the RMSEs and uncertainties are highest during May, which corresponds to melt onset in this region [8].

Table 7.1: Results Summary Between Methods for May of 2021

Method	**RMSE**	**Epistemic**	**Aleatoric**
MLP_{NT2}	0.283	N/A	N/A
BBB_{NT2}	0.260	0.066	N/A
$HBBB_{NT2}$	0.249	**0.015**	**0.166**
$CHBBB_{NT2}$	**0.235**	0.018	0.170
$CHBBB_{NT2-only-tb}$	0.301	0.008	0.190
$CHBBB_{NT2-no-tb}$	0.368	0.015	0.214

plot of Figure 7.2(d) shows some areas of the large consolidated ice patch to have small degree of error. In this area, the model predicts SIC values in an approximate range of 0.975-1.0. As the ground truth SIC for this area has a SIC of exactly 1.0, this can lead to slight deviations in the error measurement, which suggest a slight tendency of the model to under-predict. The epistemic uncertainty (Figure 7.2(f)) has 0% uncertainty in nearly all areas where the ground truth SIC is 0, such as in the Davis Strait and Northern Labrador Sea, as well as a large area in Baffin Bay by the 70th to 75th parallel.

Where the $CHBBB_{NT2}$ model predictions differ from the expected SIC is in areas near the ice edge, such as in the southern portion of Baffin Bay near Davis Strait, and the North Water Polyna, corresponding to marginal ice zones (MIZs). Epistemic uncertainty predictions in this area are relatively higher than previous, corresponding to values in the range of 3-4%. When plotting the epistemic and aleatoric uncertainties against SIC bins as seen in Figures 7.3(a) and (c), this pattern can be further illustrated. Both epistemic and aleatoric uncertainty are lowest in the SIC bins of 0 to 0.3, with increasing mean uncertainty as the SIC increases. The highest uncertainties for both are within the SIC bins of 0.6-0.9, which promptly decreases for the SIC bin of 0.9-1.0. The SIC values in the 0.9-1.0 bin have both uncertainty values higher than the values for SIC bins between 0.0-0.3, which may reflect openings in the ice cover that change day-to-day or changes in the surface conditions due to melt onset.

Observing the spatial patterns of the epistemic and aleatoric uncertainty in the $CHBBB_{NT2}$ algorithm as seen in Figure 7.2(f) and 7.2(h), both uncertainties are highest along the MIZ and ice edge. Lastly, both uncertainties have very slight signatures along the coasts, specifically in the open water region near the eastern coast of Labrador. When observing feature maps of TBH and TBV (Figures 3.3(a) and 3.3(b) respectively), the TB's have similar structure as that of the uncertainties along these coasts. This signifies land contamination from the TB data for pixels that overlap with the land-ocean boundaries.

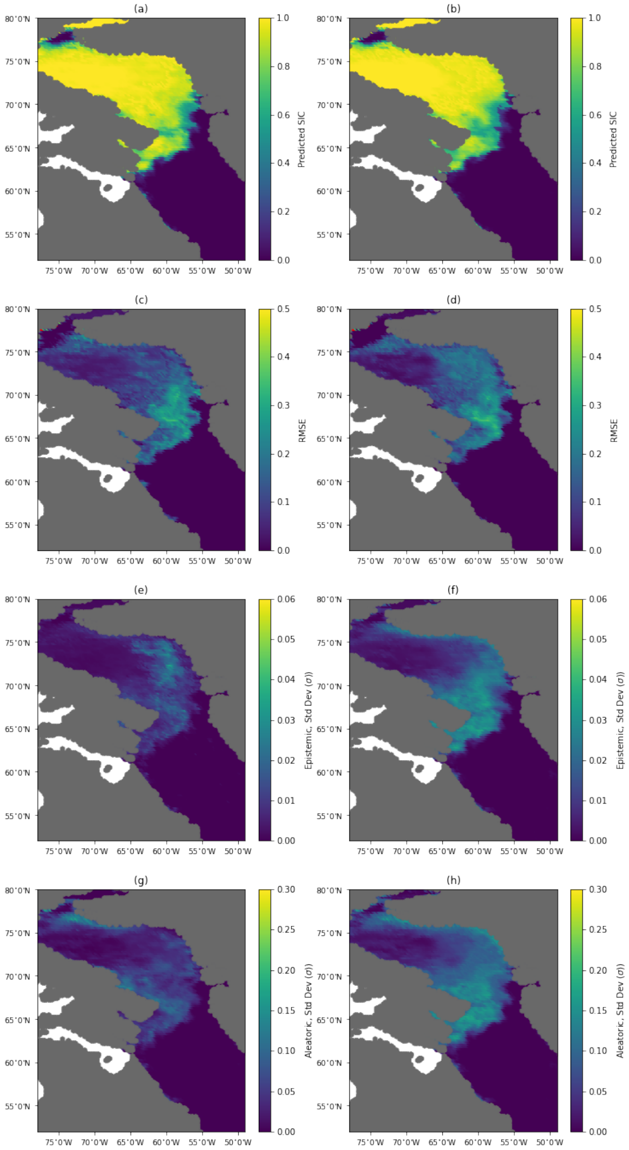

Figure 7.2: Left: $HBBB_{NT2}$ model, right: $CHBBB_{NT2}$ model. Models use 6 climate variables as input, where TB values are obtained from the AMSR2 sensor, and SIC labels used in training are calculated via the enhanced NASA team (NT2) algorithm. The calibrated model is trained to a calibration level of 0.95. Rows correspond to predicted sea ice concentrations (SIC), root mean squared error between ground truth and predicted, epistemic uncertainty, and aleatoric uncertainty in descending order from the top.

7.4 Calibration Results

In this section, we compare the results of the non-calibrated vs. calibrated models, i.e. $HBBB_{NT2}$ and $CHBBB_{NT2}$, as seen in Figure 7.2 for May of 2021. Observing the SIC predictions of both models (Figure 7.2(a),(b)), there are no significant differences. This is also reflected in the RMSE plots of each model (Figure 7.2(c),(d)). Yet, when comparing the uncertainty maps, there is a difference. The $CHBBB_{NT2}$ model can be seen to have both higher epistemic and aleatoric uncertainty as compared to that of the $HBBB_{NT2}$ model. This same behaviour can also be observed for $CHBBB_{BT}$ and $HBBB_{BT}$ models in Figure 7.5. Per Table 7.1, the differences between the $CHBBB_{NT2}$ and $HBBB_{NT2}$ models translate to an average of 2% increase in both epistemic and aleatoric uncertainty across the domain.

Furthermore, to aid in the analysis of calibration, we look to the calibration plot of Figure 7.4 between the calibrated models of $CHBBB_{NT2}$ and $CHBBB_{BT}$ and uncalibrated $HBBB_{NT2}$ and $HBBB_{BT}$ models. Here, the plot shows that both $HBBB_{NT2}$ and model $HBBB_{BT}$ over-predicts the observed frequency of ground truth SIC values at all observed confidence levels, where $HBBB_{NT2}$ over-predicts at a greater severity. For instance, at an observed confidence level of 0.2 or 20%, the calibration curve of the $HBBB_{BT}$ model shows that roughly 30% of ground truth SIC values are contained within their respective predictive intervals as produced by the predicted SIC (i.e. Equations 4.18 and 4.19), and for the $HBBB_{NT2}$ model that value is closer to 50%. Both calibrated models $CHBBB_{NT2}$ and $CHBBB_{BT}$ achieve near perfect calibration for the confidence levels of 0.0 - 0.3, with slight under predictions for higher confidence levels, although with less severity than the over prediction as seen in the $HBBB_{NT2}$ model.

7.5 Training Label Spatial Analysis

Comparing SIC prediction results and RMSE's between $CHBBB_{NT2}$ and $CHBBB_{BT}$ models (Figures. 7.2(b) and 7.5(b)), it is apparent that the $CHBBB_{BT}$ produces SIC estimates that encompass a larger spread than that of the $CHBBB_{NT2}$. This is consistent with the larger grid scale (25 km) for the SIC BT training labels, as opposed to that of the 12.5 km SIC NT2 training labels. This is most apparent near the ice edge, and the North Water Polynya, with slight differences along the Labrador Coast. The MIZ area predicted by the $CHBBB_{BT}$ model is larger than that of the $CHBBB_{NT2}$, translating to higher RMSE values in this area.

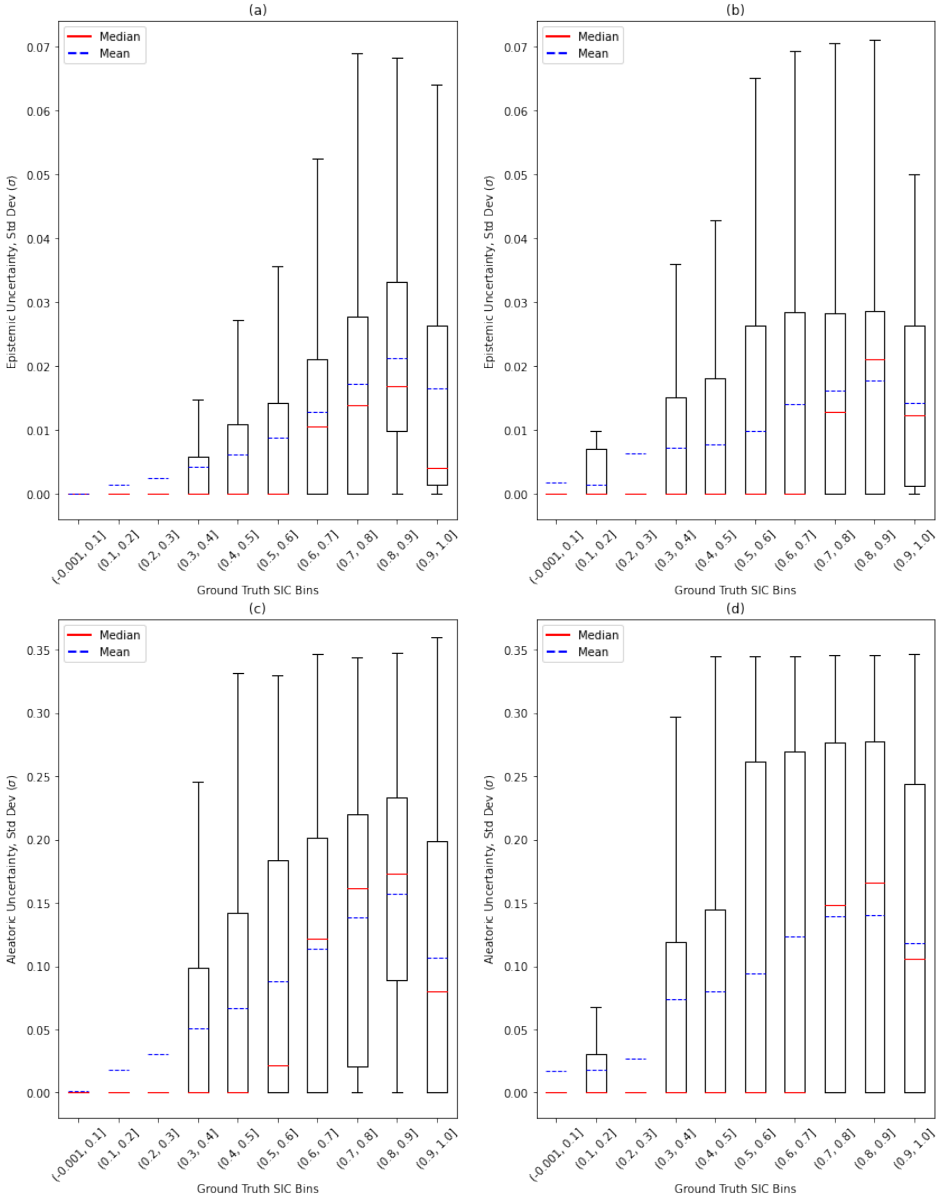

Figure 7.3: Box whisker plots of ground truth SIC vs epistemic uncertainty (top) and aleatoric uncertainty (bottom) from $CHBBB_{NT2}$ (left) and $CHBBB_{NT2}$ (right) models for May 2021. The ground truth SIC is divided into 10 equal sized bins. The box whisker plots show the mean, median, interquantile range, and extremes of uncertainties for each SIC bin.

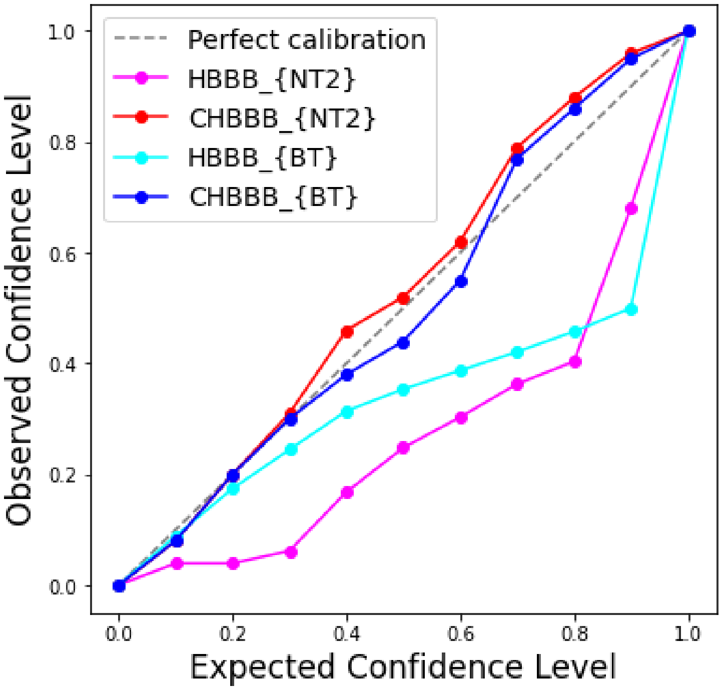

Figure 7.4: Calibration plots for the models $HBBB_{NT2}$, $CHBBB_{NT2}$, $HBBB_{BT}$, and $CHBBB_{BT}$.

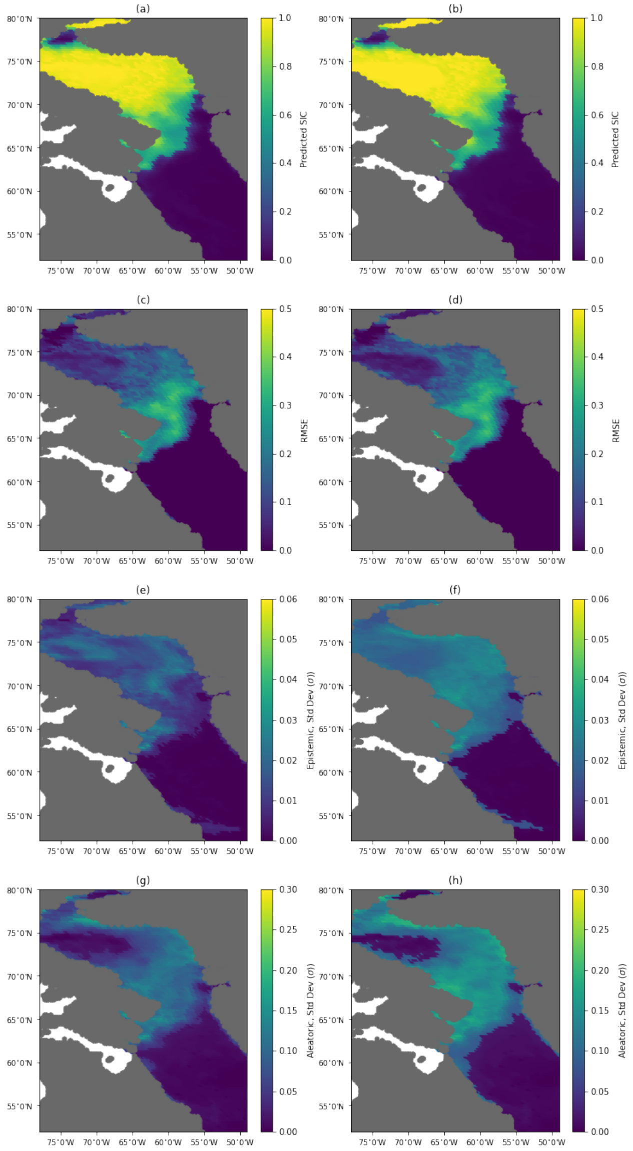

Figure 7.5: Left: $HBBB_{BT}$ model, right: $CHBBB_{BT}$ model. Models use 6 climate variables as input, where TB values are obtained from the AMSR2 sensor, and SIC labels used in training are calculated via the bootstrap (BT) algorithm. The calibrated model is trained to a calibration level of 0.95. Rows correspond to predicted sea ice concentrations (SIC), root mean squared error between ground truth and predicted, epistemic uncertainty, and aleatoric uncertainty in descending order from the top.

When observing the epistemic uncertainties of both models (Figures 7.2(f) and 7.5(f)), the epistemic uncertainty of $CHBBB_{BT}$ is in a consistent area throughout the consolidated ice region, while the $CHBBB_{NT2}$ epistemic uncertainty is localized to areas near the MIZ and ice edge. In addition, aleatoric uncertainties of each model (Figures 7.2(h) and 7.5(h)) exhibit similar patterns, however the $CHBBB_{BT}$ aleatoric uncertainty spans a slightly larger area. This area extends the uncertainty south past the ice edge, and further north towards the north water polynya. Lastly, we perceive larger signature's of both epistemic and aleatoric uncertainty for the $CHBBB_{BT}$ model in near the Hudson strait outflow and the southern open water region, specifically near the eastern coasts of Labrador and western coasts of Greenland.

When plotting the epistemic and aleatoric uncertainties for SIC bins as seen in Figure 7.3(b) and (d) for the $CHBBB_{BT}$ model, a similar pattern to that of the $CHBBB_{NT2}$ model can be seen, although with slightly different magnitudes of the mean at each bin. Here both mean epistemic and aleatoric uncertainty are lowest in the SIC bins of 0 to 0.3, and highest within the bins of 0.6-0.9.

7.6 Feature Analysis

7.6.1 Analysis on Combinations of Input Features

We first turn our attention to the spatial results from the experiments of $CHBBB_{NT2-only-tb}$. The SIC prediction of $CHBBB_{NT2-only-tb}$ (Figure 7.6(a)) shows that the model can predict the structure of the large consolidated ice covered region. However, it falters near the MIZ and open water region, where its predicts some SIC, which is originally expected to be open water. Observing the brightness temperatures, TBH and TBV, in Figures 3.3(a) & 3.3(b), we notice that these TB signatures in the open water are similar to the signatures of predicted SIC in the MIZ. The overlap in said feature space could be the reason why the model has difficulty differentiating between intermediate values of SIC and weather impacted TBs without the presence of additional atmospheric data.

Next, the epistemic uncertainty of $CHBBB_{NT2-only-tb}$ (Figure 7.6(e)) is relatively low in all areas of the region, where small sporadic signatures can be observed near the ice edge and in the open water region. Compared to the epistemic uncertainty however, the aleatoric uncertainty (Figure 7.6(g)) is consistent to the predicted SIC as seen in Figure 7.6(a), where the highest values of aleatoric uncertainty are observed to be in the consolidated ice region, and smaller signatures present in open ocean.

Following these observations from the model with TBs as the only features, we exhibit $CHBBB_{NT2-no-tb}$, the model with all features other than TBs. For this scenario, the SIC prediction (Figure 7.6(b)) shows the model's inability to capture the same sharp detail as that of $CHBBB_{NT2}$ (Fig 7.2(b)) or $CHBBB_{NT2-only-tb}$ (Fig 7.6(a)), ultimately leading to higher levels of RMSE. As per Figures 3.3(c)-(f), none of the features delineate the SIC and open water regions as clearly as the TB's (Figures 3.3(a)-(b)). This is made apparent by the low correlation these atmospheric variables have to SIC as compared to the TB's, illustrated in Figure 7.9, which may help to explain the difficulties this model has to estimate SIC. For the epistemic uncertainty of this model, values of 3-4% are consistent across most of the domain, the exception being open water, where values are slightly higher. Observing the epistemic uncertainty, it is highest within the consolidated ice region, with some small patterns again contained over open ocean. Lastly, the aleatoric uncertainty values are greatest in high SIC areas, where uncertainty is roughly in the range of 25-30%. Although both $CHBBB_{no-tb}$ and $CHBBB_{only-tb}$ models suffer from aleatoric uncertainty in most parts of the study region (Figure 7.6(g)-(h)), when all 6 features are used as input, as is the case for the $CHBBB_{NT2}$ model (Figure 7.2(f)), the aleatoric uncertainty in open water is reduced to zero, while the aleatoric uncertainty in areas of consolidated ice are then isolated to areas of the MIZ.

Additionally, for comparison, the average results of the models $CHBBB_{NT2-only-tb}$ and $CHBBB_{NT2-no-tb}$ for May of 2021 can be found in Table 7.1. It can be seen from this table that both models suffer from higher averages of RMSE, and aleatoric uncertainty for the whole study region, and $CHBBB_{NT2-no-tb}$ produces the worst performance. We note though, that both models have lower average epistemic uncertainty than the $CHBBB_{NT2}$ model.

Lastly, we explore the annual cycle with respect to the $CHBBB_{NT2-only-tb}$ model. The annual cycle of the $CHBBB_{NT2-only-tb}$ model (Figure 7.7) has a RMSE trends and magnitudes which differ significantly from the annual cycle of both the $CHBBB_{NT2}$ and $CHBBB_{BT}$ (Figure 7.1). First, the RMSE overall is much higher in magnitude. Second, the RMSE trend of the $CHBBB_{NT2-only-tb}$ model shows that the RMSE is higher for the summer months. This is directly related to TB signatures over open water, similar to the result seen in Figure 7.6(c), where large portions of open water were erroneously predicted to have SIC, resulting in higher average values of RMSE. As the summer months are characterized with larger areas of open water and less sea ice, this affect is pronounced for the whole study region. Next, epistemic uncertainties for the whole cycle are shown to have slightly smaller magnitudes in most months, while aleatoric uncertainties are observed to have higher magnitudes.

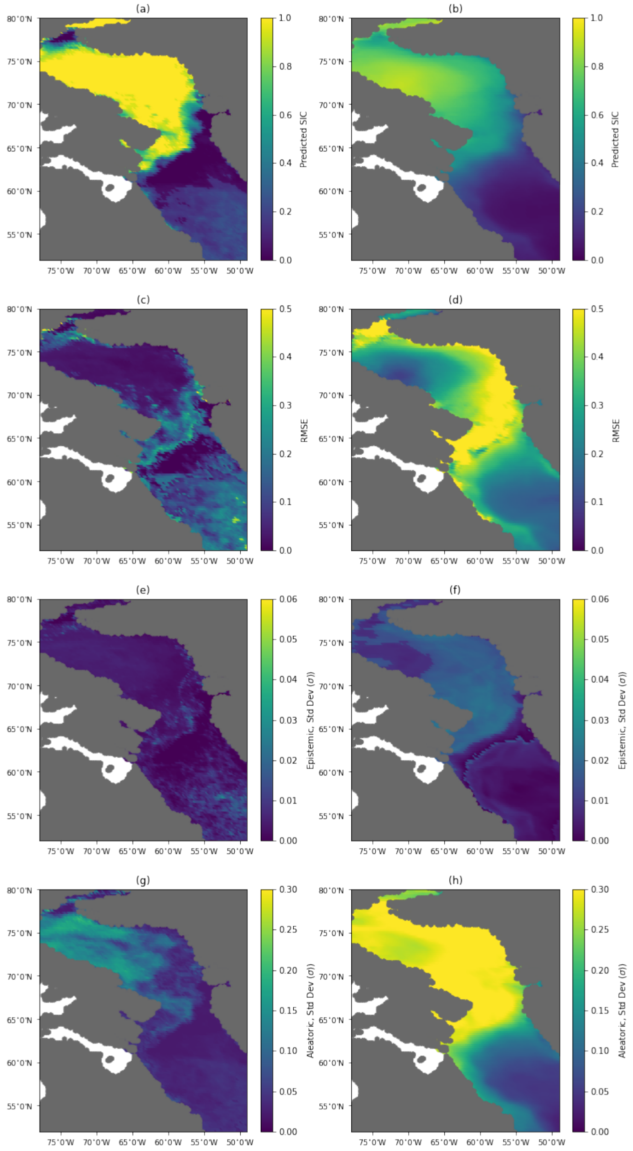

Figure 7.6: Left: $CHBBB_{NT2-only-tb}$ model, right: $CHBBB_{NT2-no-tb}$, Model $CHBBB_{NT2-only-tb}$ uses only TB's from AMSR2 as input, while $CHBBB_{NT2-no-tb}$ uses the 4 atmospheric climate variables of WS, WV, CW, and AT. Both use SIC labels calculated from the NT2 algorithm, and are trained until a calibration level of 0.95 is reached. Rows correspond to predicted SIC, RMSE between ground truth and predicted, epistemic uncertainty, and aleatoric uncertainty in descending order from the top.

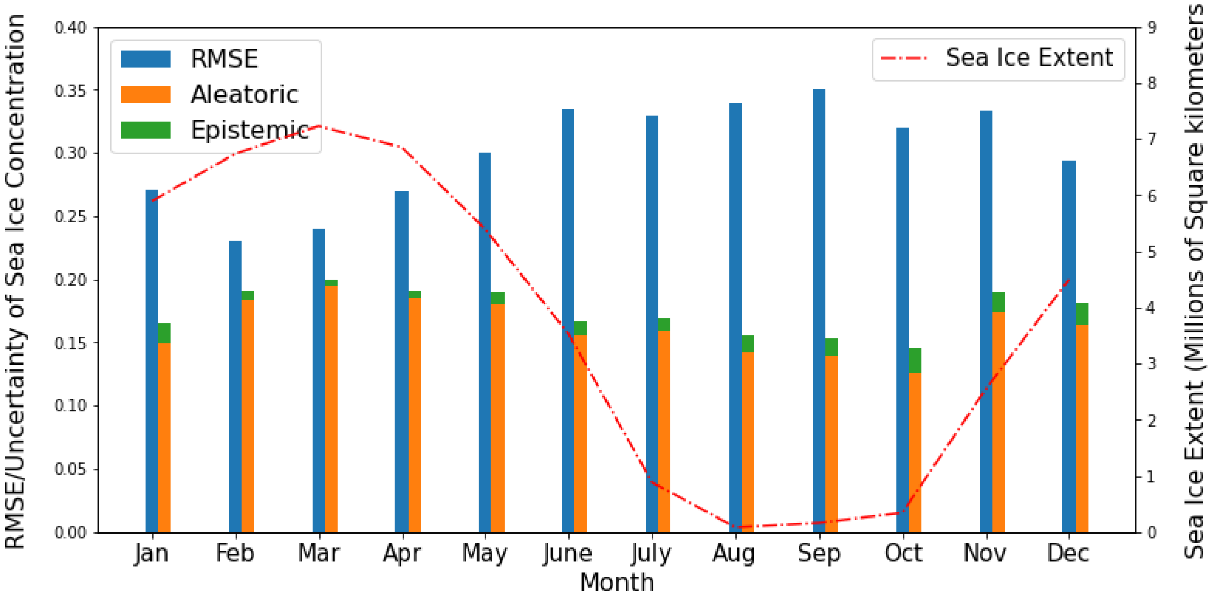

Figure 7.7: Root mean squared error (RMSE), epistemic uncertainty, aleatoric uncertainty using the $CHBBB_{NT2-only-tb}$ model. The aleatoric and epistemic uncertainty are shown as a stacked bar, representing total uncertainty. The sea ice extent (SIE) values averaged per month for the testing of year of 2021 are also shown for reference to the seasonal cycle. Note the RMSEs and uncertainties are highest during May, which corresponds to melt onset in this region [8].

7.6.2 Analysis on input features

Moreover, we look at specific details of the features as they pertain to the annual cycle. The heatmaps of TBH and TBV (Figures 7.8(a)-(b)), show that all features have highest cumulative uncertainties in the month of May. WS values of 2-5 m/s are shown to have the highest total uncertainty, while WV has the highest total uncertainties for values between 5-10 mm. The final two features of CW and AT have high cumulative uncertainty in bins of 0-0.1 mm and 268-275 K respectively. These observation suggest that values in these ranges for each feature contribute in some way to conditions that favor high uncertainty. Relating this back to the monthly observations of uncertainty and RMSE in Section 7.1, these conditions are ones that favor rapid rates of melt specifically in this region.

In addition, all features, except CW, have some distribution of uncertainty spread to various bins across months. Differing from the others, cloud water has the highest uncertainty in the bins of 0 to 0.1 mm regardless of the month. When analyzing the correlation matrix between features used in $NT2$ models (Figure 7.9), CW has the lowest correlation to SIC as compared to the other features, with a value of -0.23. There are more moderate negative correlations of SIC between WS and WV, with a strong negative correlation to AT. The feature with the highest correlation to SIC are both TB's, where both TB's also show strong positive correlations between one another. The TB's have a similar correlation relationship to the other 4 features of WS, WV, CW, and AT, but with the lowest correlation to CW. The same results were found for the correlation matrix of features used in BT models (Figure 7.10)

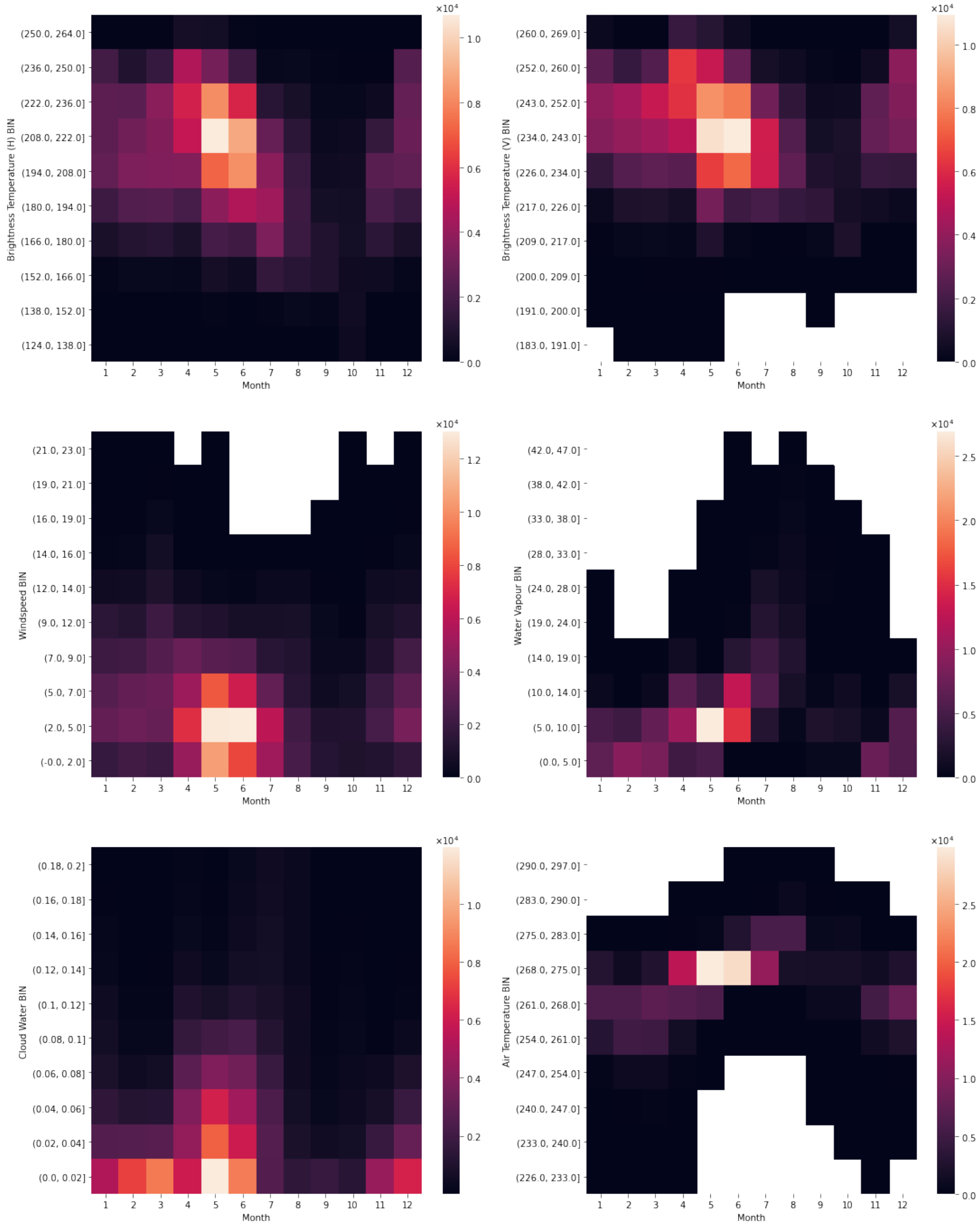

Figure 7.8: Heatmaps of input features, plotted with bins of their respective values against month. The colorbars represent cumulative total uncertainty as outputted by the $CHBBB_{NT2}$ model. For most features, May distinctly shows the highest cumulative uncertainty.

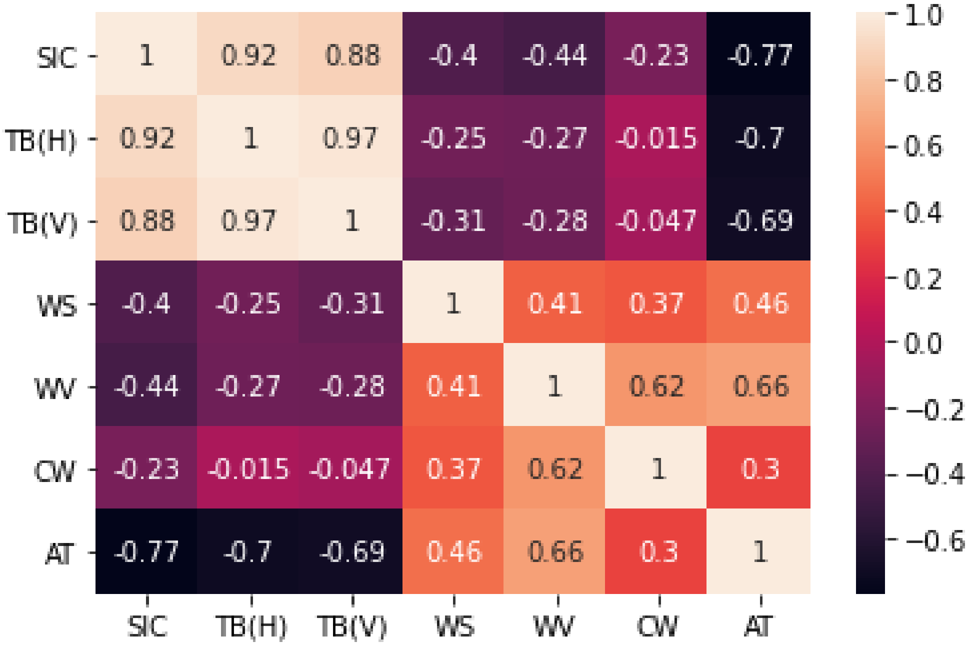

Figure 7.9: Correlation matrix between features and SIC. TB values are from the AMSR2 sensor, and SIC training labels are calculated from the NT2 algorithm.

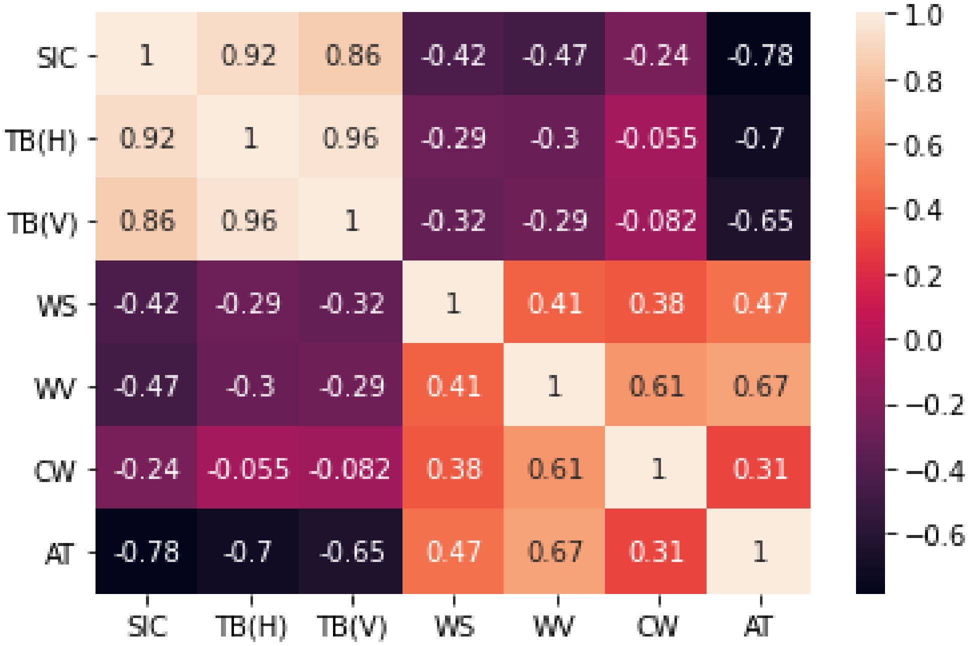

Figure 7.10: Correlation matrix between features and SIC. TB values are from the SSMIS sensor, and SIC training labels are calculated from the BT algorithm.

Chapter 8

Concluding Remarks

8.1 Discussion

When analyzing predictions of SIC from $CHBBB_{NT2}$ and $CHBBB_{BT}$ over the annual cycle of 2021, the highest values for both RMSE and uncertainty are shown to be in the months of April and May, where May coincides with the melt onset in this region [8]. Within this period of melt onset and the following period of summer, SIC retrieval algorithms have high uncertainty, and estimates of SIC can be poor [55], [38]. Additionally, for months with greater variability in SIC, there can be larger differences in retrievals of SIC, due to greater influences by atmospheric effects, and surface conditions [68]. To the contrary, the lowest values of RMSE and uncertainty coincide with times in the annual cycle corresponding to low SIC. This result was found without the need for weather filters or TB corrections within the methodology, which had been used in previous studies [26], [13], [75], [4], [3]. The model had also shown promise to predict SIC in times of freeze up, such as in the months of November and December. These trends suggests that the uncertainty varies from season to season, consistent with the findings of Brucker et. al. [10], which found that uncertainty of NT2 varied seasonally as well. Prediction of uncertainties as a function of SIC peaks for higher values of intermediate ice concentration. Similar results were found in the study by Tonboe et. al. [75] based on tie-point standard deviations and smearing due to the large footprint associated with the PM measurements. The advantage of the approach presented in this study though is its data-driven nature. Given new changes to the data, spatially or temporally, uncertainty estimates are specific to that data instance. Along side these benefits, are the production of spatial uncertainty maps.

Subsequently, when comparing between methodologies, it was shown that the use of the heteroscedastic loss as opposed to homoscedastic loss (MSE loss) as likelihood in our models produce lower epistemic (model) uncertainty as well as greater accuracy in SIC predictions. The aleatoric uncertainty changes slightly between methodologies, but from Kendall et. al [41], these changes in magnitude were minimal, which is consistent with our findings. As noted later, these changes are a directly result of the calibration methodology. Though when exploring experiments done on different subsets of input features, i.e. $CHBBB_{NT2-no-tb}$, and $CHBBB_{NT2-only-tb}$, it was found that when utilizing the full set of input features, as in the case for $CHBBB_{NT2-no-tb}$, we observed the reduction of aleatoric uncertainty. This suggests the addition of features to the feature space, as opposed to adding more data, help reduce the aleatoric uncertainty. Methods such as Monte Carlo dropout for measuring epistemic uncertainty were explored in past trials but did not show capabilities to effectively predict SIC and estimate epistemic uncertainty for these predictions. Ensemble methods were also considered, but required greater computational costs then that of a Bayesian neural network, and due to limited computational power were not used.

Additionally, from spatial observations of the $CHBBB_{NT2}$ model, it has been observed that the model had larger uncertainty for SIC value in the MIZ and along the sea ice edge. In general, these large uncertainties are captured in ice covered regions, as opposed to open water regions, similar to the another method of PM based SIC uncertainty estimation [10]. We also looked at the differences between the $CHBBB_{NT2}$ and $CHBBB_{BT}$ models, where it was first observed from the monthly observations a difference in the RMSE and uncertainties between the models (Figure 7.1). When comparing the spatial behaviour of $CHBBB_{BT}$ to the that of the $CHBBB_{NT2}$ estimates, the former slightly underestimates SIC in regions of the MIZ as opposed to the latter. Furthermore, epistemic and aleatoric uncertainty for the $CHBBB_{BT}$ model span larger areas, which include some pronounced areas along the coasts in the domain. The differences between the two algorithms may be attributed to the spatial resolution of each SIC retrieval algorithm, which stems from the TBs used by each algorithm. Higher resolution data are able to introduce finer detail during training, and as a result, when performing predictions, the results using the higher resolution data are less diffuse than that of the results using lower resolution data, leading to differences in SIC estimates, and their respective RMSEs and uncertainties. Irregardless of these differences, choice of training label do not pose a challenge to estimate SIC, which shows the flexibility of the methodology to garner accurate predictions.

Next, when examining the the calibrated models of $CHBBB_{NT2}$ and $CHBBB_{BT}$, it was observed that both had higher values of epistemic and aleatoric uncertainty than that of the uncalibrated models $HBBB_{NT2}$ and $HBBB_{BT}$. From Equation 4.18, the approximate predictive interval (PI) has two ways of adjusting itself, either through the aleatoric standard deviation (uncertainty) $\hat{\sigma_a}$ or mean prediction $\hat{y}$. It may be possible that both are adjusted and accounted for in the model, but from Figure 7.2 of $CHBBB_{NT2}$ and $HBBB_{NT2}$, as well for $CHBBB_{BT}$ and $HBBB_{BT}$ (Figure 7.5) the predicted mean does not appear to have changed as significantly as compared to the aleatoric uncertainty. The changes to the aleatoric uncertainty are apparent as there is an increase in the aleatoric uncertainty of the calibrated model(s) as compared to the uncalibrated model(s). We also observe that the epistemic uncertainty between the calibrated model(s) compared to the uncalibrated model(s) increases as well. We argue that this effect is caused by the adjustment of the predicted mean. As the epistemic uncertainty is the resulting deviations between different samples as drawn from the model, this suggests that single samples of SIC mean predictions in the calibrated model have changed such that they are more variable than that of the uncalibrated model, resulting in higher measures of epistemic uncertainty. Additionally, this may also be attributed to the bias-variance trade off [43], where as the model increase in complexity, i.e. transforming non-calibrated models to calibrated models which require more complex model architecture, the model variance is expected to increase. Thereafter, although the model produces higher estimates of uncertainty, we reason that these estimates provide better quality and produce a representation aligned with the definition of calibration. As a result, both the $CHBBB_{NT2}$ and $CHBBB_{BT}$ are significantly more calibrated than their respective uncalibrated counterparts of $HBBB_{NT2}$ and $HBBB_{BT}$ which is seen from the calibration plot of Figure 7.4. Additionally, both calibration curves of $CHBBB_{NT2}$ and $CHBBB_{BT}$ are in close agreement with one another, which indicates that the methodology can produce calibrated models, and reduces the dependency on the choice of SIC used as a training label. Recall though that both models are trained to a calibration level of 0.95. We would expect that all observed confidence levels would match perfectly to the expected confidence level for all observed confidence levels until 0.95. This is not the case however, as the models are over confident at higher confidence levels, implying further training or hyper parameter tuning is needed.

Finally, from the analysis on input feature combinations, a few things come to light. The features of brightness temperature contribute the most to producing sharp and detailed SIC predictions as seen in the maps of predicted SIC (7.6(a)-(b)). This is exemplified through the high correlation between TBs and SIC, as shown by the correlation matrices in Figures 7.9 and 7.10. The TBs, combined with the 4 features of wind speed, column water vapour, liquid water, and air temperature are able to help correct the over prediction of $CHBBB_{NT2-only-tb}$ in open water regions, as reflected in the RMSE maps in Figures 7.2(c)-(d) 7.6(c)-(d). Next, the models which used less features ($CHBBB_{NT2-no-tb}$ and $CHBBB_{NT2-only-tb}$), contributed less epistemic uncertainty in most areas of the study region as compared to the $CHBBB_{NT2}$ model. This may be attributed to the bias-variance trade off [43], where increasing model complexity (i.e. adding more features) has been known to increase the variance. This variance as described by the bias-variance trade off differs from the epistemic uncertainty, where it has been suggested that the epistemic uncertainty is actually a sum of the bias and variance [24]. This may suggest that the epistemic uncertainty for the $CHBBB_{NT2}$ model has a larger contribution from the model variance then the bias as compared to the $CHBBB_{NT2-only-tb}$ and $CHBBB_{NT2-no-tb}$ models, but requires further study on epistemic uncertainty bias-variance decomposition. Though the model suggests an increase in variance, the trade off though, is the greater accuracy (i.e. lower RMSE's) in predicting SIC when using all 6 input features, as is the case for the $CHBBB_{NT2}$ model. Finally, The combination of features also reduces the expected aleatoric uncertainty of the model for $CHBBB_{NT2}$ (Figure 7.2(h)), localizing it to areas of consolidated ice. This does not necessarily affect the bias-variance trade off, given that the aleatoric uncertainty is a measure of the noise as opposed to model variance. Overall, the combination of all 6 features evidently helps in improvements of SIC estimation, reduction of RMSE, and reduction of aleatoric uncertainty.

8.2 Conclusion

The study as shown in this thesis has conveyed the capabilities of a heteroscedastic Bayesian MLP model calibrated with auxiliary interval predictors using PM-TB and atmospheric reanalysis data to produce accurate predictions of SIC, and the quantification of both epistemic (model) and aleatoric (data) uncertainty. The model has shown the ability to evaluate SIC over the full annual cycle of 2021 in a seasonal ice zone, including times of melt, freeze up, and solid ice. The model had also exhibited pockets of high uncertainty in the study region, such as in the marginal ice zone (MIZ), and along the ice edge.

We observed that as calibration was performed, there was an overall increase of uncertainty. Since calibration is the chosen method in deep learning to evaluate the uncertainties produced by these models, we argue that producing a calibrated model produces greater quality, and trustworthiness in uncertainty estimates as opposed to the uncalibrated model.

Finally, the choice of features has shown to affect the prediction of SIC, epistemic, and aleatoric uncertainty. When using solely brightness temperatures (TB) as an input feature, it was to capture spatial extent of SIC well, due to their high correlation and the use of TB in SIC algorithms, but produced erroneous predictions of SIC and high uncertainty in open water due to noise. The 4 atmospheric climate variables windspeed, air temperature, cloud water vapour, and cloud liquid water, could not capture spatial extent well, but when combined with TBs, they aid in reducing epistemic and aleatoric uncertainties, especially over open water.

8.3 Future Work

Given the expansive research in the fields of machine learning and deep learning as applied to remote sensing, as well as growing data availability, there are several aspects that can be improved upon in this study for future work.

With the continuous rise of data content in sea ice remote sensing it may be of interest in the future to encompass decades worth of data to use for training and inference. We have observed the uncertainty (and error) of the model to vary seasonally, with highest uncertainties coinciding with times of melt onset. If a larger dataset is to be introduced, which is not limited to computational resources (and time), the model may learn to generalize better and produce greater accuracy in mean predictions of SIC, but also reduce the model uncertainty in these high uncertainty periods. To add, it may also be feasible to incorporate a larger dataset which reach greater domains, such as that of the whole arctic. We have observed the model to produce varying predictions spatially, where highest uncertainties coincide with areas of the MIZ. Similar to the reasoning for introducing yearly data, introducing a larger region scope may help to reduce uncertainty in these high uncertainty regions, while also encapsulating regions of uncertainty for the whole arctic. For example, regions known to have a significant fraction of melt pond coverage. If other data such as independent in-situ data such as from SAR or optical imagery are available, it may be of interest to compare these trained models to this independent data to measure the accuracy of our predictions.

Next, the ERA5 dataset contains a plethora of atmospheric and oceanic variables. We have seen with the addition of 4 atmospheric climate variables the improvements of the model as opposed to the input of only using brightness temperatures. A research direction could be to utilize a larger quantity of such variables as input into the model, while observing the uncertainties that originate from this data. This may help to develop greater accuracy of SIC predictions while also decreasing uncertainties. It is also of great interest to explore new calibration techniques for regression, to compare to the methods used in this study, and if greater computational resources are available, to utilize alternative methodologies in uncertainty quantification, such as that of deep ensembles, as well as epistemic uncertainty decomposition to measure the effects of bias and variance, which may help lead suggestions on model complexity and reducing epistemic uncertainty.

References

[1] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021.

[2] Leonidas Alagialoglou, Ioannis Manakos, Marco Heurich, Jaroslav x010C;ervenka, and Anastasios Delopoulos. A learnable model with calibrated uncertainty quantification for estimating canopy height from spaceborne sequential imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022.

[3] Søren Andersen, Rasmus Tonboe, Lars Kaleschke, Georg Heygster, and Leif Toudal Pedersen. Intercomparison of passive microwave sea ice concentration retrievals over the high-concentration Arctic sea ice. *Journal of Geophysical Research: Oceans*, 112, 2007.

[4] Søren Andersen, Rasmus Tonboe, Stefan Kern, and Harald Schyberg. Improved retrieval of sea ice total concentration from spaceborne passive microwave observations using numerical weather prediction model fields: An intercomparison of nine algorithms. *Remote Sensing of Environment*, 104:374–392, 10 2006.

[5] Tom Andersson, J. Hosking, María Pérez-Ortiz, Brooks Paige, Andrew Elliott, Chris Russell, Stephen Law, Dani Jones, Jeremy Wilkinson, Tony Phillips, Steffen Tietsche, Beena Sarojini, Eduardo Blanchard-Wrigglesworth, Yevgeny Aksenov, Rod Downie, and Emily Shuckburgh. Seasonal Arctic sea ice forecasting with probabilistic deep learning. 02 2021.

[6] Nazanin Asadi, K. Andrea Scott, Alexander S. Komarov, Mark Buehner, and David A. Clausi. Evaluation of a neural network with uncertainty for detection of ice and water

in SAR imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 59(1):247–259, 2021.

[7] H. Bi, Z. Zhang, Y. Wang, X. Xu, Y. Liang, J. Huang, Y. Liu, and M. Fu. Baffin Bay sea ice inflow and outflow: 1978–1979 to 2016–2017. *The Cryosphere*, 13(3):1025–1042, 2019.

[8] Angela Bliss, Michael Steele, Ge Peng, W. Meier, and Suzanne Dickinson. Regional variability of Arctic sea ice seasonal change climate indicators from a passive microwave climate data record. *Environmental Research Letters*, 14, 04 2019.

[9] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. 2015.

[10] Ludovic Brucker, Donald J. Cavalieri, Thorsten Markus, and Alvaro Ivanoff. NASA team 2 sea ice concentration algorithm retrieval uncertainty. *IEEE Transactions on Geoscience and Remote Sensing*, 52(11):7336–7352, 2014.

[11] Frank D. Carsey. *Microwave Remote Sensing of Sea Ice, Volume 68.* Geophysical Monograph Seriess. American Geophysical Union, 1992.

[12] 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.

[13] Donald J. Cavalieri, Karen M. St. Germain, and Calvin T. Swift. Reduction of weather effects in the calculation of sea-ice concentration with the DMSP SSM/I. *Journal of Glaciology*, 41(139):455–464, 1995.

[14] Junhwa Chi and Hyun-choel Kim. Prediction of Arctic sea ice concentration using a fully data driven deep neural network. *Remote Sensing*, 9(12), 2017.

[15] J.C. Comiso. *Bootstrap sea ice Concentrations from Nimbus-7 SMMR and DMSP SSM/I-SSMIS, Version 3.* NASA National Snow and Ice Data Center Distributed Active Archive Center.

[16] 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.

[17] J.C. Comiso and F. Nishio. *Enhanced Sea Ice Concentrations from Passive Microwave Data, Bootstrap Algorithm), date = 2008, organization = NASA Goddard Space Flight Center,.*

[18] Josefino Comiso and Fumihiko Nishio. Trends in the sea ice cover using enhanced and compatible AMSR-E, SSM/I, and SMMR data. *Journal of Geophysical Research*, 113, 05 2007.

[19] Colin L. V. Cooke and K. Andrea Scott. Estimating sea ice concentration from SAR: Training convolutional neural networks with passive microwave data. *IEEE Transactions on Geoscience and Remote Sensing*, 57(7):4735–4747, 2019.

[20] Liuhui Ding, Dachuan Li, Bowen Liu, Wenxing Lan, Bing Bai, Qi Hao, Weipeng Cao, and Ke Pei. Capture uncertainties in deep neural networks for safe operation of autonomous driving vehicles. *CoRR*, abs/2108.05118, 2021.

[21] Moira Dunbar. The geographical position of the North Water. *Arctic*, 22(4):438–441, 1969.

[22] Alexandre Gagnon and William Gough. Trends in the dates of ice freeze-up and breakup over Hudson Bay, Canada. *Arctic*, 58:370–382, 12 2005.

[23] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. 2015.

[24] Jing Gao. Bias-variance decomposition of absolute errors for diagnosing regression models of continuous data. *Patterns*, 2(8):100309, 2021.

[25] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna M. Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. A survey of uncertainty in deep neural networks. *CoRR*, abs/2107.03342, 2021.

[26] P. Gloersen and D. J. Cavalieri. Reduction of weather effects in the calculation of sea ice concentration from microwave radiances. 91(C3):3913–3919, March 1986.

[27] Irina Gorodetskaya, Bruno Tremblay, Beate Liepert, Mark Cane, and Richard Cullather. The influence of cloud and surface properties on the Arctic ocean shortwave radiation budget in coupled models. *Journal of Climate*, 21, 03 2008.

[28] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. 2017.

[29] Jarrod Haas and Bernhard Rabus. Uncertainty estimation for deep learning-based segmentation of roads in synthetic aperture radar imagery. *Remote Sensing*, 13(8), 2021.

[30] W.D. Halliday, J. Dawson, D.J. Yurkowski, T. Doniol-Valcroze, S.H. Ferguson, C. Gjerdrum, N.E. Hussey, Z. Kochanowicz, M.L. Mallory, M. Marcoux, C.A. Watt, and S.N.P. Wong. Vessel risks to marine wildlife in the Tallurutiup Imanga national marine conservation area and the eastern entrance to the Northwest Passage. *Environmenal Science and Policy*, 127:181–192, 2022.

[31] Hyangsun Han and Hyuncheol Kim. Evaluation of summer passive microwave sea ice concentrations in the Chukchi Sea based on KOMPSAT-5 SAR and numerical weather prediction data. *Remote Sensing of Environment*, 209:343–362, 2018.

[32] Hyangsun Han, Sungjae Lee, Hyun-Cheol Kim, and Miae Kim. Retrieval of summer sea ice concentration in the Pacific Arctic Ocean from AMSR2 observations and numerical weather data using random forest regression. *Remote Sensing*, 13(12), 2021.

[33] Ligong Han, Yang Zou, Ruijiang Gao, Lezi Wang, and Dimitris N. Metaxas. Unsupervised domain adaptation via calibrating uncertainties. *CoRR*, abs/1907.11202, 2019.

[34] H. Hersbach, B. Bell, P. Berrisford, G. Biavati, A. Horányi, J. Muñoz Sabater, J. Nicolas, C. Peubey, I. Radu, R. amd Rozum, D. Schepers, A. Simmons, C. Soci, D. Dee, and J-N. Thépaut. *(2018): ERA5 hourly data on single levels from 1979 to present.* Copernicus Climate Change Service (C3S) Climate Data Store (CDS).

[35] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991.

[36] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

[37] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110, 03 2021.

[38] N. Ivanova, L. T. Pedersen, R. T. Tonboe, S. Kern, G. Heygster, T. Lavergne, A. Sørensen, R. Saldo, G. Dybkjær, L. Brucker, and M. Shokr. Inter-comparison and evaluation of sea ice algorithms: towards further identification of challenges and

optimal approach using passive microwave observations. *The Cryosphere*, 9(5):1797–1817, 2015.

[39] Yuchen Jiang, Shen Yin, and Okyay Kaynak. Data-driven monitoring and safety control of industrial cyber-physical systems: Basics and beyond. *IEEE Access*, 6:47374–47384, 2018.

[40] James Joyce. Bayes' Theorem. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2021 edition, 2021.

[41] Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? *CoRR*, abs/1703.04977, 2017.

[42] Jiwon Kim, Kwangjin Kim, Jaeil Cho, Yong Q. Kang, Hong-Joo Yoon, and Yang-Won Lee. Satellite-based prediction of Arctic sea ice concentration using a deep neural network with multi-model ensemble. *Remote Sensing*, 11(1), 2018.

[43] Ron Kohavi and David Wolpert. Bias plus variance decomposition for zero-one loss functions. In *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*, ICML'96, page 275–283. Morgan Kaufmann Publishers Inc., 1996.

[44] Benjamin Kompa, Jasper Snoek, and Andrew L Beam. Communicating uncertainty in medical machine learning. *npj Digital Medicine*, 4(1):1–6, 2021.

[45] Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. *CoRR*, abs/1807.00263, 2018.

[46] C. Kummerow. On the accuracy of the Eddington approximation for radiative transfer in the microwave frequencies. *Journal of Geophysical Research: Atmospheres*, 98(D2):2757–2765, 1993.

[47] K.L. Laidre, H. Stern, K.M. Kovacs, L. Lowry, S.E. Moore, Regehr E.V., S.H. Ferguson, Ø. Wiig, R.P. Boveng, P.and Angliss, E.W. Born, L. Litovka, D.and Quakenbush, C. Lydersen, D. Vongraven, and F. Ugarte. Arctic marine mammal population status, sea ice habitat loss, and conservation recommendations for the 21st century. *Conservation Biology*, 29:724–737, 2015.

[48] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. NIPS'17, page 6405–6416. Curran Associates Inc., 2017.

[49] Quoc V. Le, Alex J. Smola, and Stéphane Canu. Heteroscedastic gaussian process regression. ICML '05, page 489–496, New York, NY, USA, 2005. Association for Computing Machinery.

[50] Yann LeCun, Y. Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–44, 05 2015.

[51] Matti Leppranta. *Drift ice material*, pages 11–63. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[52] Dan Levi, Liran Gispan, Niv Giladi, and Ethan Fetaya. Evaluating and calibrating uncertainty prediction in regression tasks. *CoRR*, abs/1905.11659, 2019.

[53] Gongbo Liang, Yu Zhang, Xiaoqin Wang, and Nathan Jacobs. Improved trainable calibration method for neural networks on medical imaging classification. *CoRR*, abs/2009.04057, 2020.

[54] T. Markus, D.J. Cavalieri, and J.C. Comiso. *Algorithm Theoretical Basis Document Sea Ice Products*. NASA Goddard Space Flight Center.

[55] W. Meier and D. Notz. A note on the accuracy and reliability of satellite-derived passive microwave estimates of sea-ice extent. Technical report, CLIC International Project Office, 2010.

[56] W. N. Meier, T. Markus, and J. C. Comiso. *AMSRE AMSR2 unified L3 daily 25 km brightness temperatures, sea ice concentration, motion snow depth polar grids, Version 1.* NASA National Snow and Ice Data Center Distributed Active Archive Center.

[57] W. N. Meier, J. S. Stewart, H. Wilcox, D.J. Scott, and H.A. Hardman. *DMSP SSM/I-SSMIS daily polar gridded brightness temperatures, Version 6.* NASA National Snow and Ice Data Center Distributed Active Archive Center.

[58] C. J. Merchant, F. Paul, T. Popp, M. Ablain, S. Bontemps, P. Defourny, R. Hollmann, T. Lavergne, A. Laeng, G. de Leeuw, J. Mittaz, C. Poulsen, A. C. Povey, M. Reuter, S. Sathyendranath, S. Sandven, V. F. Sofieva, and W. Wagner. Uncertainty information in climate data records from Earth observation. *Earth System Science Data*, 9(2):511–527, 2017.

[59] P.J. Minnett, A. Alvera-Azcárate, T.M. Chin, G.K. Corlett, C.L. Gentemann, I. Karagali, X. Li, A. Marsouin, S. Marullo, E. Maturi, R. Santoleri, S. Saux Picart, M. Steele,

and J. Vazquez-Cuervo. Half a century of satellite remote sensing of sea-surface temperature. *Remote Sensing of Environment*, 233:111366, 2019.

[60] P.J. Minnett and E.L. Key. Chapter 4 meteorology and atmosphere–surface coupling in and around Polynyas. In W.O. Smith and D.G. Barber, editors, *Polynyas: Windows to the World*, volume 74 of *Elsevier Oceanography Series*, pages 127–161. Elsevier, 2007.

[61] Jeremy Nixon, Mike Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. *CoRR*, abs/1904.01685, 2019.

[62] Manfred Opper and Cedric Archambeau. The variational gaussian approximation revisited. *Neural computation*, 21:786–92, 10 2008.

[63] Pedro Ortiz, Marko Orescanin, Veljko Petkovi;, Scott W. Powell, and Benjamin Marsh. Decomposing satellite-based classification uncertainties in large earth science datasets. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022.

[64] Liang Peng, Hong Wang, and Jun Li. Uncertainty evaluation of object detection algorithms for autonomous vehicles. *Automotive Innovation*, 4, 07 2021.

[65] L. Pizzolato, S.E.L. Howell, J. Dawson, F. Laliberté, and L. Copland. The influence of declining sea ice on shipping activity in the Canadian Arctic. *Geophysical Research Letters*, 43:12,146–12,154, 2015.

[66] Keerthijan Radhakrishnan, K. Scott, and David Clausi. Sea ice concentration estimation: Using passive microwave and SAR data with a u-net and curriculum learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, PP:1–1, 04 2021.

[67] Sivaramakrishnan Rajaraman, Prasanth Ganesan, and Sameer Antani. Deep learning model calibration for improving performance in class-imbalanced medical image classification tasks. *PLOS ONE*, 17, 01 2022.

[68] Anja Rosel and Lars Kaleschke. Exceptional melt pond occurrence in the years 2007 and 2011 on the Arctic sea ice revealed from MODIS satellite data. *Journal of Geophysical Research: Oceans*, 117(C5).

[69] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

[70] Nastaran Saberi, Katharine Andrea Scott, and Claude Duguay. Incorporating aleatoric uncertainties in lake ice mapping using RADARSAT-2 SAR images and CNNs. *Remote Sensing*, 14(3), 2022.

[71] Armina Soleymani and K. Andrea Scott. Evaluation of a neural network on sea ice concentration estimation in MIZ using passive microwave data. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 5656–5659, 2021.

[72] 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.

[73] Jayaraman J. Thiagarajan, Kowshik Thopalli, Deepta Rajan, and Pavan Turaga. Training calibration-based counterfactual explainers for deep learning models in medical image analysis. 12(1), 1 2022.

[74] Jayaraman J. Thiagarajan, Bindya Venkatesh, Prasanna Sattigeri, and Peer-Timo Bremer. Building calibrated deep models via uncertainty matching with auxiliary interval predictors. 2019.

[75] 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.

[76] F.T. Ulaby, Moore R.K., and A.K. Fung. *Microwave Remote Sensing - Active and Passive*. Artech House, 1981.

[77] Lei Wang, K. Scott, Linlin Xu, and David 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*, 54:1–10, 04 2016.

[78] Lei Wang, K. Andrea Scott, and David A. Clausi. Sea ice concentration estimation during freeze-up from SAR imagery using a convolutional neural network. *Remote Sensing*, 9(5), 2017.

[79] P.D. Wasserman and T. Schwartz. Neural networks. ii. what are they and why is everybody so interested in them now? *IEEE Expert*, 3(1):10–15, 1988.