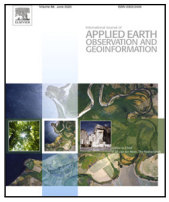
International Journal of Applied Earth Observation and Geoinformation 131 (2024) 103920

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag



A comparative study of data input selection for deep learning-based automated sea ice mapping

Xinwei Chen [a], Fernando J. Pena Cantu [b], Muhammed Patel [b], Linlin Xu [b,*], Neil C. Brubacher [b], K. Andrea Scott [c], David A. Clausi [b]

[a] School of Marine Science and Engineering, South China University of Technology, 777 Xingyedadao East Rd, Guangzhou, 511400, Guangdong, China
[b] Vision and Image Processing (VIP) Research Group, Department of Systems Design Engineering, University of Waterloo, 200 University Ave W, Waterloo, N2L 3G1, ON, Canada
[c] Department of Mechanical and Mechatronics Engineering, University of Waterloo, 200 University Ave W, Waterloo, N2L 3G1, ON, Canada

ARTICLE INFO

Dataset link: https://data.dtu.dk/collections/AI4Arctic_Sea_Ice_Challenge_Dataset/6244065/2, https://github.com/echonax07/MMSeaIce

Keywords:
Sea ice Mapping
Synthetic aperture radar (SAR)
AMSR2
Multi-source Data
Feature importance
AI4Arctic

ABSTRACT

The precise monitoring of sea ice parameters, including sea ice concentration and stage of development, is imperative for tactical navigation. Recent studies have showcased the enhanced mapping accuracy achieved by incorporating multi-source auxiliary data, such as passive microwave data, with Synthetic Aperture Radar (SAR) images. However, there remains a lack of research assessing the impact of individual features on model performance. This paper addresses this knowledge gap through ablation studies and alternate comparisons of data inputs. Building on the success in the AutoIce Challenge, we leverage the AI4Arctic Sea Ice Challenge Dataset to train multitask sea ice mapping models employing a U-Net architecture. Results from cross-validation and testing sets with all season data reveal the significant enhancement in estimation accuracy for all parameters when utilizing most of the AMSR2 channels. Additionally, the incorporation of time and location information as ancillary channels further amplifies the classification accuracy of all major ice types. Furthermore, among the various available ERA5 weather parameters, the inclusion of wind speed data proves effective in mitigating misclassifications in ice regions, particularly under melting scenarios. The paper culminates with a feature importance ranking table encompassing all available features, providing valuable guidance for the selection of pertinent data inputs. This comprehensive comparative study not only contributes to advancing sea ice mapping methodologies but also offers valuable insights into the nuanced impact of individual features on model performance.

1. Introduction

Sea ice is a crucial component in earth systems model for climate projections (Holland and Bitz, 2003; Day et al., 2022). The substantial reduction in Arctic sea ice extent and thickness over the past decades, driven by Arctic warming, has garnered global attention. Simultaneously, the decline in Arctic sea ice has opened up previously impassable Arctic sea routes, sparking commercial interests in establishing more direct connections among global markets (Mudryk et al., 2021).

Efficient maritime shipping in cold and remote polar regions necessitates real-time information about local sea ice conditions for safe and swift navigation. Traditionally, this information is conveyed through sea ice charts, manually generated by trained ice analysts at national ice services. These charts rely on visual interpretation of Synthetic Aperture Radar (SAR) imagery, ancillary multi-source data, and information from icebreakers (Sandven et al., 2023). However, this manual process is time-consuming, exhibits limited spatial and temporal coverage, and is susceptible to operator biases (Tamber et al., 2022; Chen et al., 2024b). The thinning of Arctic sea ice, coupled with an increasing volume of satellite data and wider accessibility to Arctic regions, intensifies the challenges of traditional manual ice charting.

The potential automation of the labor-intensive sea ice charting process holds promise for delivering near-real-time sea ice products with higher spatial resolution, and increased temporal coverage. Operational sea ice mapping involves estimating three crucial parameters: sea ice concentration (SIC), stage of development (SOD), and floe size (FLOE). SAR, with its high spatial resolution, polarimetric capability, and flexible imaging modes, stands out as a preferred solution for detailed sea ice mapping (Li et al., 2021). Recent years have witnessed the application of machine learning (ML) and deep learning (DL) techniques, especially convolutional neural networks (CNNs), for automated sea ice

* Corresponding author.
E-mail address: l44xu@uwaterloo.ca (L. Xu).

https://doi.org/10.1016/j.jag.2024.103920
Received 29 January 2024; Received in revised form 7 May 2024; Accepted 16 May 2024
Available online 8 June 2024

Table 1
The features (data inputs) available in the dataset, which are all utilized for investigation in this research.

Data input type	Data input (feature) name	Spatial resolution	Total number of channels
SAR variables	HH	80 × 80 m	3
	HV		
	Incidence angle (IA)		
Passive microwave (AMSR2 brightness temperature) data in horizontal and vertical polarizations	AMSR2 6.9 GHz	2 × 2 km	14
	AMSR2 7.3 GHz		
	AMSR2 10.7 GHz		
	AMSR2 18.7 GHz		
	AMSR2 23.8 GHz		
	AMSR2 36.5 GHz		
	AMSR2 89.0 GHz		
ERA5 weather data	Eastward component of 10 m wind speed (u10)		6
	Northward component of 10 m wind speed (v10)		
	Temperature 2 m above the surface (T2M)		
	Skin temperature (SKT)		
	Total water column water vapor (TCWV)		
	Total column cloud liquid water (TCLW)		
Ancillary data	Distance-to-land map	80 × 80 m	4
	Time information: acquisition month	N/A	
	Location information: SAR latitude grid	~19 × 19 km	
	Location information: SAR longitude grid		

parameter estimation from dual-polarized SAR imagery (Wang et al., 2016; Boulze et al., 2020; Nagi et al., 2021; De Gelis et al., 2021; Stokholm et al., 2022; Ren et al., 2021).

Despite the successes of DL approaches, ambiguities persist in the relationship between SAR backscatter and ice conditions. Distinct ice types, concentrations, and various wind conditions may yield similar backscatter signatures (Malmgren-Hansen et al., 2020; Radhakrishnan et al., 2021). Solely relying on SAR imagery as input for DL models might lead to inaccurate mapping results. To address this challenge, researchers have explored multi-sensor approaches, combining the strengths of different techniques to mitigate ambiguities (Gabarró et al., 2023). Examples include the fusion of Sentinel-1 SAR imagery and AMSR2 passive microwave data for SIC estimation (Malmgren-Hansen et al., 2020; Karvonen, 2017) and SOD prediction using heterogeneous data (SAR and optical images) (Han et al., 2021b). In addition to multi-sensor approaches, ancillary information can be included for performance improvement. For instance, in Park et al. (2020), SAR data were separated into multiple groups based on acquisition month/season to train separate sea ice classifiers. The AI4Arctic Sea Ice Challenge dataset (Buus-Hinkler et al., 2022), released in 2022, serves as a valuable resource for advancing DL-based mapping of multiple sea ice parameters. The dataset includes Sentinel-1 SAR images, Advanced Microwave Scanning Radiometer 2 (AMSR2) brightness temperature data, numerical weather prediction data from ECMWF Reanalysis v5 (ERA5), and other ancillary data. Given the wealth of satellite and ancillary data, understanding the efficacy of various data inputs in improving model performance is critical. However, a comprehensive investigation into the importance of different data inputs in multiple sea ice mapping tasks is lacking.

This study expands the work done on our previous preliminary works (Chen et al., 2023, 2024a), by conducting an in-depth analysis of the various data inputs and their contributions to sea ice mapping. Employing ablation studies and statistical significance testing, we determine which features are crucial for enhancing sea ice mapping performance, which have minimal impact, and which are redundant. Unlike our previous works (Chen et al., 2023, 2024a), this analysis focuses on understanding the impact of different features on the performance of key sea ice parameters: SOD, SIC, FLOE, along with their respective subclasses. We discuss the reasons behind the varying influences of these features on specific classes. Finally, based on our findings, we provide well-informed recommendations for selecting the most effective features when developing DL models for sea ice mapping.

An introduction of the study area and the multi-source AI4Arctic dataset is given in Section 2. The methodology, including the CNN architecture and the experiments concerning feature selection process, is illustrated in Section 3. Section 4 provides the sea ice mapping results using various combinations of data inputs, along with comparisons and discussions. In the end, the conclusions summarizing the findings of the research, as well as future works, are given in Section 5.

2. Data

The AI4Arctic Sea Ice Challenge Dataset, unveiled in 2022, is a comprehensive resource comprising 533 files, distributed between the training dataset (513 files) and the held-out testing dataset (20 files). The dataset encompasses two distinct versions: a raw version and a ready-to-train version. For the purposes of this study, we utilized the ready-to-train version, which is pre-processed and optimized for DL model implementations, as detailed in the data manual (Buus-Hinkler et al., 2022). Each file in the dataset encapsulates a diverse array of multi-source data, as outlined in Table 1.

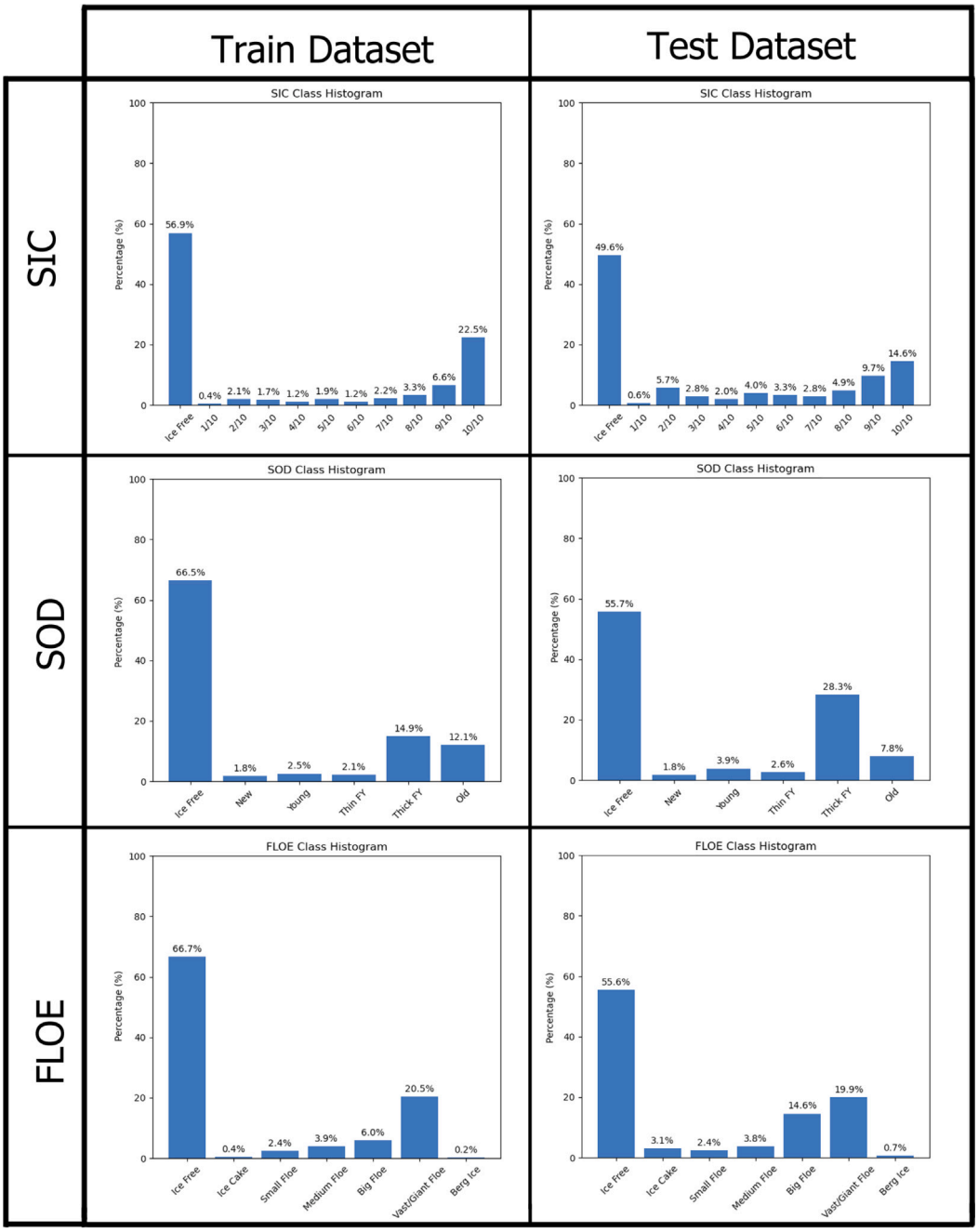


Fig. 1. Class distribution between training and testing dataset for SIC, SOD, and FLOE.

The dataset incorporates Sentinel-1 Extra Wide Swath (EW) SAR scenes, offering data in both HH and HV polarizations. Notably, these SAR scenes have undergone noise correction using the NERSC algorithm (Korosov et al., 2021). To balance computational efficiency and high-resolution mapping, a 2 × 2 kernel has been applied to downsample the pixel spacing to 80 m (~5000 × 5000 pixels) in the ready-to-train version.

In tandem with SAR data, the dataset includes corresponding AMSR2 brightness temperatures, resampled to the Sentinel-1 geometry with a 2 km grid. Simultaneously, ERA5 weather parameters are retrieved and resampled to align with the Sentinel-1 geometry. Finally, supplementary ancillary data, detailed at the bottom of Table 1, is added to further enrich the dataset. Information about the spatial resolutions of those data is given in Table 1.

The data is then normalized to fall within the $[-1, 1]$ range, achieved by scaling to the minimum and maximum values for each variable, in order to facilitate the model learning. Data collection spans the period from January 2018 to December 2021, encompassing the Arctic regions surrounding Canada and Greenland.

Pixel-based labels are vital for model training and evaluation. In the ready-to-train version, these labels are derived from polygon-based egg codes found in ice charts. Specifically, the SIC attributed to the egg code of each polygon is assigned to all pixels within the polygon as ground truths. Beyond total SIC, each polygon contributes partial SICs, each associated with a specific SOD and FLOE. Three partial concentrations, summing to the total SIC, offer a nuanced perspective. The SOD and FLOE labels for pixels within a polygon are determined by the dominant ice type and its corresponding FLOE, defined as the ice type with at least 65% in the partial concentration normalized by the total SIC.

Table 2
Layer specifications of the U-Net-based model using all available data inputs.

	Name	Input feature maps	Output feature maps
Encoder	Convolution block 1	27 × 256 × 256	32 × 256 × 256
	Max-pooling 1	32 × 256 × 256	32 × 128 × 128
	Convolution block 2	32 × 128 × 128	32 × 128 × 128
	Max-pooling 2	32 × 128 × 128	32 × 64 × 64
	Convolution block 3	32 × 64 × 64	64 × 64 × 64
	Max-pooling 3	64 × 64 × 64	64 × 32 × 32
	Convolution block 4	64 × 32 × 32	64 × 32 × 32
	Max-pooling 4	64 × 32 × 32	64 × 16 × 16
	Bridge	64 × 16 × 16	64 × 16 × 16
Decoder	Up-sample 1	64 × 16 × 16	64 × 32 × 32
	Convolution block 5	128 × 32 × 32	64 × 32 × 32
	Up-sample 2	64 × 32 × 32	64 × 64 × 64
	Convolution block 6	128 × 64 × 64	64 × 64 × 64
	Up-sample 3	64 × 64 × 64	64 × 128 × 128
	Convolution block 7	96 × 128 × 128	32 × 128 × 128
	Up-sample 4	32 × 128 × 128	32 × 256 × 256
	Convolution block 8	64 × 256 × 256	32 × 256 × 256
Outputs	1 × 1 convolution+ Softmax (SIC)	32 × 256 × 256	12 × 256 × 256
	1 × 1 convolution+ Softmax (SOD)	32 × 256 × 256	7 × 256 × 256
	1 × 1 convolution+ Softmax (FLOE)	32 × 256 × 256	8 × 256 × 256

Polygons lacking a dominant ice type are excluded from both training and evaluation processes.

The distribution of classes for SIC, SOD, and FLOE in training and testing are shown in 1. The distribution is heavily unbalance with most of the pixels belonging to open water, and some of the classes having less than 1% of total pixels.

3. Methodology

3.1. Data input processing

Utilizing the diverse features as inputs for the CNN-based sea ice mapping model involves a structured three-step process, as illustrated in Fig. 2. In Chen et al. (2024a), we built a multi-task U-Net for extracting SIC, SOD, and FLOE using the AI4Arctic dataset. Building on insights from this prior work, we emphasize the pivotal role of downsampling SAR data to significantly enhance mapping accuracy by facilitating the model in capturing contextual information. In the first step, the extensive $\sim 5,000 \times 5,000$ pixel SAR images, alongside the distance-to-land map and corresponding pixel-based label maps, undergo a downsampling by a factor of 10. This results in a more manageable size of approximately $\sim 500 \times 500$ pixels with a pixel spacing of around 800 m. Subsequently, the data in subgridded variables (comprising all other variables specified in Table 1) are then upsampled to match the resolution of the SAR/label map. This strategic alignment enables the concatenation of all features into a cohesive set serving as data inputs for the model. Finally, to ensure uniformity in data inputs during model training, random cropping is applied to the data inputs using a patch size of 256 × 256. A meticulous grid search, encompassing the selection of downsampling rate and patch size, has been conducted. The values outlined above have been identified as optimal, yielding the most robust and effective model performance.

3.2. U-Net-based sea ice mapping model

Building on our successful participation in the AutoIce sea ice mapping challenge (Stokholm et al., 2023; Chen et al., 2024a), we continue to leverage the effectiveness of the U-Net architecture (Ronneberger et al., 2015) for our sea ice mapping model. As depicted in Fig. 2 and detailed in Table 2, the adopted U-Net architecture comprises four encoder and decoder blocks seamlessly connected by a bridge layer. Within each convolution block, two convolutional layers are employed, with the number of filters specified by the digits in Fig. 2. The final up-sample layer's output is intricately linked to three distinct 1 × 1 convolution layers. This connection serves the crucial purpose of generating mapping results for the three sea ice parameters.

Table 3
Specifications of model training.

Learning rate	0.0001
Weight decay	0.001
momentum	0.001
Optimizer	SGD
Batch size	16
Number of iterations per epoch	500
Total epoch	60
Patch size	256
Down Sample	10x
Loss functions	Cross entropy

3.3. Training specifications

Model training specifications are elucidated in Table 3. Optimal hyperparameters, crucial for achieving peak performance, are diligently determined through an exhaustive grid search. To guarantee ample exposure per data sample during the training phase, each epoch encompasses 500 iterations. Within each iteration, a batch of patches is randomly extracted from the training scenes. During these iterations, the cross-entropy losses computed for the three parameters are aggregated, collectively constituting the total loss. This strategic amalgamation of losses ensures a comprehensive optimization approach, considering all three sea ice parameters.

3.4. Model evaluation metrics

To assess the model's performance, we employ the metrics established in the AutoIce Challenge (Stokholm et al., 2023). For SIC,

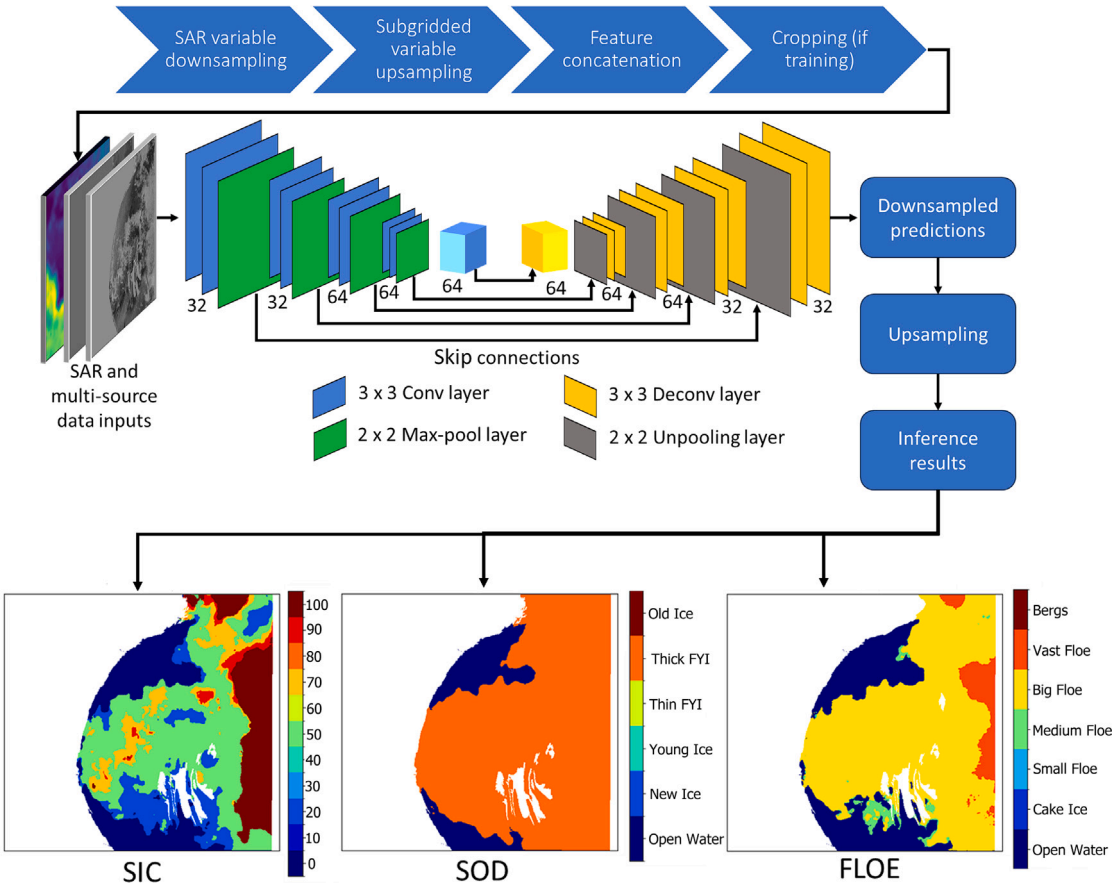


Fig. 2. A diagram illustrating the U-Net-based sea ice mapping pipeline implemented in this research.

accuracy is quantified using the R2 score. This score captures the inter-class relationships among different SIC levels, acknowledging that, for instance, 10% SIC is closer to 20% than to 30%. Mathematically, the R2 score is defined as:

$$\text{R2 Score} = 1 - \frac{\sum_{i=1}^{N_{\text{pixel}}}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N_{\text{pixel}}}(y_i - \bar{y}_i)^2}, \tag{1}$$

where y_i and $\hat{y}_i$ represent the ground truth and estimated SIC of the ith pixel, respectively. $\bar{y}$ denotes the mean value of all y_i among all valid pixels in all SAR scenes of the validation/testing set, and N_{pixel} is the total number of such pixels.

For SOD and FLOE, the F1 score is chosen as the evaluation metric:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{2}$$

with Precision and Recall defined as:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}, \tag{3}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}, \tag{4}$$

where true positives, false positives, and false negatives are aggregated across all valid pixels in the validation/testing set.

To offer a consolidated assessment, a combined score is introduced in Stokholm et al. (2023), defined as:

$$\text{Combined Score} = \frac{2}{5} \times \text{SIC} + \frac{2}{5} \times \text{SOD} + \frac{1}{5} \times \text{FLOE}. \tag{5}$$

The assigned weights to the three parameters ensure that the combined score does not exceed 100%. Given the operational significance of SIC and SOD over FLOE, a relatively smaller weight is assigned to the latter.

3.5. Cross-validation and hypothesis testing

To scrutinize the influence of distinct data types on model efficacy, we embarked on a dual-faceted investigative approach, encompassing both removal and addition studies. In the removal study, we systematically eliminated specific features, such as AMSR2 channels, from the comprehensive "Full Model" (which incorporates all data types) to assess the consequential impact on model performance. Conversely, the addition study is initiated with a foundational model, exclusively employing HH and HV SAR polarization features, to which we added other data types to evaluate performance enhancements.

Ensuring the veracity of our findings, each experiment was subjected to Monte Carlo cross-validation. This entailed the random selection of 20 scenes for validation, with the remaining 493 scenes allocated for model training. Post-training, model performance was then evaluated. This procedure was replicated over 30 iterations, with results subsequently averaged to mitigate variance. Note that the test scenes do not change during this procedure.

To ascertain the statistical significance of removing/adding specific feature types on model scores, z-tests (Gupta and Guttman, 2014) were conducted between base models (models in the first rows of Tables 6 and 7) and other models. Assuming that the given sample mean score (including combined scores and individual parameter scores) is approximately normally distributed (due to the central limit theorem) and the t-test approximately equal to the z-test when sample size is larger than 30, the following hypotheses were formulated:

- **Null Hypothesis** (H_0): There is no significant difference between the two scores.
- **Alternative Hypothesis** (H_1): A certain score in the base model is significantly larger than that score in another model (or vice versa). A one-tailed test was chosen for this analysis.

Table 4
Overall performance of the "Full Model".

Model	Cross validation dataset				Test dataset				Sample Number
	Combined Score	SIC	SOD	FLOE	Combined Score	SIC	SOD	FLOE	
Full Model	91.13	92.21	92.01	87.24	83.93	87.03	86.26	73.09	34

Table 5
Monthly performance for the "Full Model". Bold highlights low scores.

Months	Combine Score	Combine Score Std	SIC	SIC Std	SOD	SOD Std	FLOE	FLOE Std	Sample number
January	**86.95**	10.32	94.58	4.68	**82.98**	18.34	**79.60**	23.69	45
February	90.51	10.28	96.30	3.52	86.82	17.46	86.33	16.40	44
March	91.63	6.12	96.26	2.43	88.53	11.92	88.56	9.22	32
April	91.17	6.15	94.24	4.35	91.60	9.33	84.19	13.64	47
May	90.75	6.68	91.84	6.05	94.65	7.92	80.79	23.01	44
June	88.90	7.35	88.42	9.58	93.84	6.21	80.00	19.11	37
July	**85.12**	17.65	**86.89**	27.84	**86.95**	13.36	**77.95**	22.29	61
August	90.66	9.34	92.26	8.13	91.25	11.27	86.25	14.70	97
September	93.19	7.81	95.01	6.17	93.44	8.62	89.04	17.84	82
October	91.54	10.29	93.64	8.03	89.76	14.94	90.89	12.03	67
November	88.68	12.34	93.69	8.74	88.30	15.22	**79.44**	26.25	45
December	**85.98**	11.29	94.45	4.46	**79.11**	22.76	82.79	15.79	53

With a significance level α of 0.05, Z-Statistics were calculated using the formula:

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \tag{6}$$

where $\bar{X}_1$ and $\bar{X}_2$ are the sample mean scores, s_1 and s_2 are the sample standard deviations, and n_1 and n_2 are the number of samples for each model. For a one-tailed test, the critical value (1.645) was obtained from the standard normal distribution table. If the calculated Z-statistic falls in the critical region, the null hypothesis is rejected.

4. Results and discussion

4.1. Performance of full model

This section elaborates on the performance of the "Full Model", which integrates all available data types for sea ice mapping. According to Table 4, the model demonstrates disparate performance across the cross-validation and test datasets, with the latter presenting a more challenging scenario as evidenced by lower average scores. The model exhibits commendable accuracy in estimating SIC and SOD, albeit encountering difficulties in accurately predicting floe sizes, which tend to score lower.

Fig. 3 shows the confusion matrixes for validation and testing and each metric (SOD, FLOE and SIC). Overall, the confusion matrixes show that the model particularly struggles with intermediate classes. In terms of SIC, it often misclassifies categories that are neither 0% nor 100%, yet tends to predict values close to the actual class, reflecting a generally satisfactory performance. For SOD, the model encountered significant difficulties with new ice, young ice, and thin FYI. In the domain of floe size, it struggles with most categories, with ice cake and icebergs presenting the most substantial challenges. The observed challenges are likely linked to the limited availability of data for certain ice types. As shown in Fig. 1 new ice, young ice, thin FYI ice, ice cake, and iceberg are relatively rare occurrences in the dataset, which may contribute to the model's underperformance in these categories.

Finally, Table 5 presents the model's monthly performance. The results were derived from over 30 Monte Carlo cross-validation runs. On each run the scores per scenes were calculated. Afterwards the scores were average across each month. The analysis reveals that the model performs optimally during transitional periods of ice melting and freezing, particularly in March and September when ice extents are at their maximum and minimum, respectively. Conversely, the model's performance dips in July, January, and December, coinciding with middle melting and freezing periods. A closer look at specific metrics indicates a significant drop in SIC accuracy only during the month of July, while SOD and floe size predictions are notably less accurate in December, January, and July. Indicating that SIC prediction is not notably worse during freeze-up.

4.2. Results removing and adding data inputs

The results of removing and adding data inputs are shown in Tables 6 and 7. Based on the z-test results, scores significantly lower than the "Full Model" in Table 6 and scores significantly higher than the "HH, HV only" model in Table 7 are bold. The ablation studies show that removing passive microwave (AMSR2) channels or time and location channels significantly impacts model scores in both cross-validation and testing sets. For instance, removing all AMSR2 channels reduces combined scores by 1.01% and 3.52% in cross-validation and testing sets, respectively, primarily due to drops in SOD scores (e.g., a 9.03% drop in the testing set). Similarly, removing time and location information channels reduces model accuracy, especially for SOD (9.04% drop in the testing set). Conversely, removing other features (e.g., ERA5, incidence angle, and distance-to-land map) from the Full model does not show statistical significance in affecting model accuracy. These findings are further validated in "adding" experiments detailed in Table 7. For instance, adding AMSR2 channels improves scores for all three parameters (2.31% and 5.52% increase in combined scores for cross-validation and testing sets, respectively). Incorporating time and location features also contributes significantly to model performance, with a 3.54% and 3.13% increase in combined scores for cross-validation and testing sets, respectively. Although adding other types of features might improve model scores, these improvements are not consistent between cross-validation and testing sets. For instance, adding ERA5 features improves the combined testing score by 3.67% but shows limited enhancement in cross-validation sets. A detailed analysis for each feature type is provided below. Furthermore, It seems ERA5 data is redundant once AMSR2 is added to model as there is no drop in performance when ERA5 is removed.

4.3. Passive microwave (AMSR2) feature analysis

An example SAR scene in the test data set with the corresponding labels and mapping results is presented in Fig. 4 to demonstrate the effectiveness of AMSR2 inputs in improving the classification accuracy of SOD. Without AMSR2 inputs, the model misclassifies the region

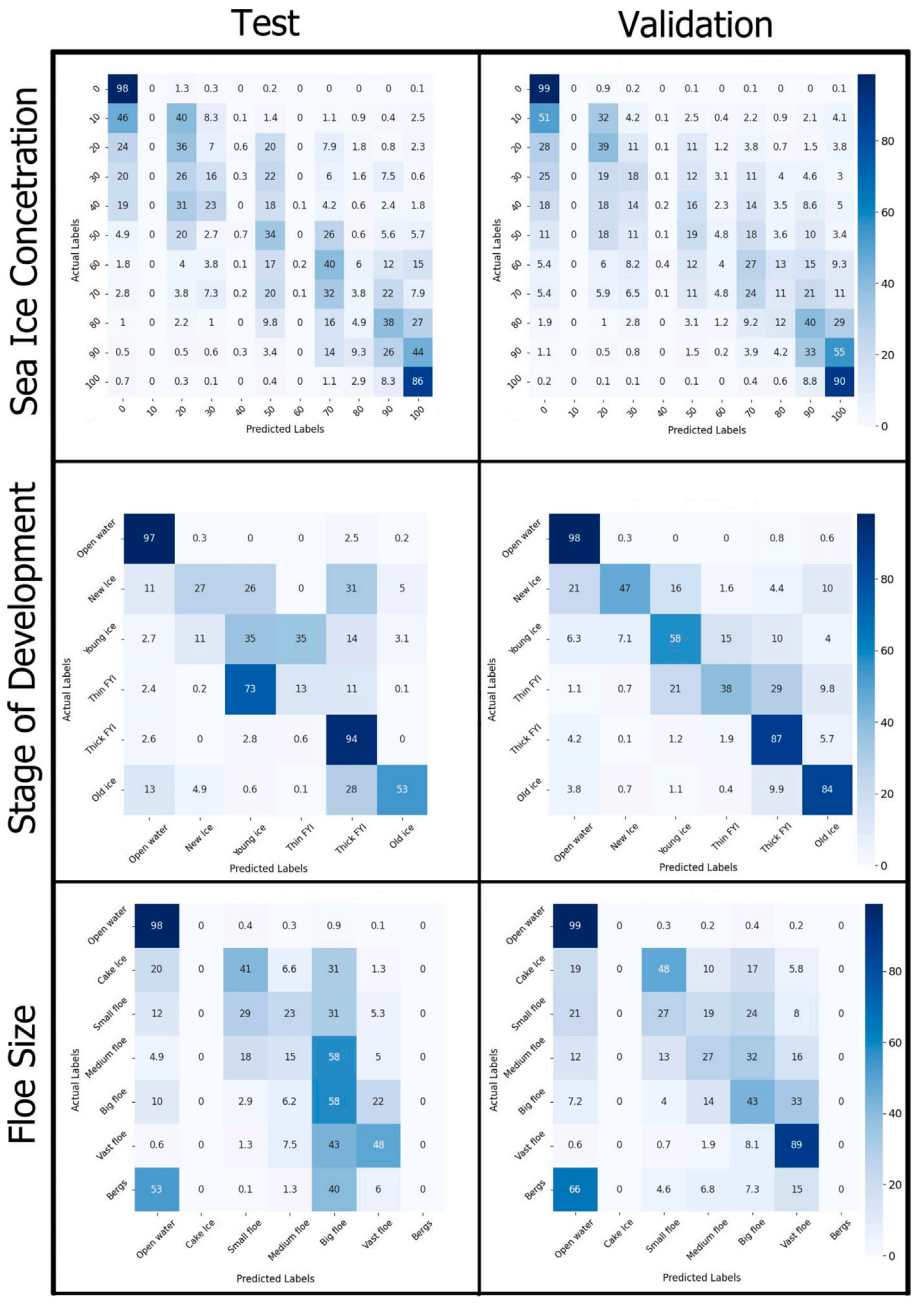


Fig. 3. Confusion matrices for SIC, SOD, and FLOE on the test and validation dataset.

with young ice as thick first-year ice (FYI). In contrast, this issue is significantly alleviated with the incorporation of AMSR2 (i.e., the full model). The improvement in SIC and FLOE is less significant in this scene.

To analyze the difference between different AMSR2 frequencies on model performance, models with each individual AMSR2 frequency channel (concatenated with SAR HH and HV channels) are trained and compared with each other, as shown in Table 8. The improvements in model performance are close for all frequencies except 89.0 GHz. Particularly, the model with AMSR2 6.9 GHz channels obtains the highest combined and SOD scores. Although the 6.9 GHz channels have relatively low resolution, they are less sensitive to noise and only sensitive to sea surface temperature as shown in Fig. 5. This could explain why they achieved the highest improvement in mapping accuracy between different frequencies. Conversely, the 89.0 GHz channels are more sensitive to weather-induced noise, causing ambiguities between ice and water.

One may hypothesis that the lower resolution of 6.9 GHz may hinder the model performance by causing poorly delineated edges. However qualitatively, these does not seem to be the case. Fig. 6 compares the output of HH + HV + AMSR2 6.9 GHz against HH + HV + AMSR2 89.0 GHz which has a much higher resolution. The scene scores for 6.9 GHz and 89.0 GHz, respectively are SIC: 93.88%, SOD: 91.52%, FLOE: 69.48% and SIC: 93.55%, SOD: 74.11%, FLOE: 69.77%. Both models tend to have boundaries between water and ice that are quite similar to each other. Furthermore, the SIC and FLOE scores are quite similar on both models. However, in this scene for SOD, 6.9 GHz significantly outperforms 89.0 GHz. The main reason being that 89.0 misclassifies entire regions as old ice. It also misclassifies the ice in the fjord as new ice instead of young ice. This indicate that the main difference in performance does not come from edges but rather misclassification of entire regions.

Additionally in Fig. 7, we can see that the channel 89.0 GHz (V) struggles to differentiate between water and ice. These two factors

Table 6
Summary of numerical results from ablation studies on different feature types. The scores in SIC, SOD, and FLOE correspond to the R2, F1, and F1 scores, respectively. Data inputs that have a statistical significant impact on model performance are in bold.

Model description	Cross validation (%)				Testing (%)			
	Combine Score	SIC	SOD	FLOE	Combined score	SIC	SOD	FLOE
Full model (No removal)	**91.13**	**92.21**	**92.01**	**87.24**	**83.93**	**87.03**	**86.26**	**73.09**
Remove IA	91.81	92.41	92.71	88.80	84.06	87.33	86.52	72.60
Remove distance map	92.13	93.07	93.02	88.49	83.46	86.89	85.52	72.50
Remove AMSR2	**90.12 (−1.01)**	92.01	**89.46 (−2.54)**	87.65	**80.41 (−3.52)**	87.12	**77.23 (−9.03)**	73.33
Remove ERA 5	92.15	92.64	93.31	88.82	84.10	87.44	86.02	73.59
Remove IA, distance map, ERA 5	91.60	92.29	92.70	88.02	83.55	86.66	85.84	72.76
Remove Time	91.76	93.17	91.92	88.61	**82.49 (−1.45)**	87.10	**82.81 (−3.46)**	72.62
Remove Location	90.78	92.62	90.98	86.71	**82.21 (−1.72)**	87.08	**81.64 (−4.63)**	73.61
Remove Time and location	**90.21 (−0.92)**	91.95	**89.93 (−2.07)**	87.32	**80.45 (−3.48)**	87.23	**77.22 (−9.04)**	73.37
Remove SAR variables (HH, HV, IA)	**89.91 (−1.22)**	**89.21 (−2.99)**	91.60	87.90	**78.51 (−5.42)**	**78.45 (−8.57)**	**82.76 (−3.50)**	**70.13 (−2.96)**

Table 7
Summary of numerical results obtained from the “adding features" experiments. Data inputs that have a statistical significant impact on model performance are in bold.

Model description	Cross validation (%)				Testing (%)			
	Combined score	SIC	SOD	FLOE	Combined score	SIC	SOD	FLOE
HH, HV only	88.01	90.75	86.10	86.37	75.63	77.58	76.19	70.60
HH, HV, and ERA 5	88.17	91.05	86.20	86.37	**79.30 (+3.67)**	**83.89 (+6.31)**	**77.58 (+1.39)**	**73.54 (+2.95)**
HH, HV, and distance map	**89.36 (+1.34)**	**91.49 (+0.75)**	**87.75 (+1.65)**	**88.29 (+1.92)**	75.20	76.69	75.90	**70.86 (0.29)**
HH, HV, and IA	**88.79 (+0.78)**	**91.44 (+0.70)**	86.96	87.15	75.81	77.79	76.56	70.37
HH, HV, and AMSR2	**90.32** (+2.31)	**92.30** (+1.55)	**90.01 (+3.91)**	87.00	**81.15 (+5.52)**	**87.28 (+9.70)**	**78.43 (+2.24)**	**74.33 (+3.73)**
HH, HV, and Time	**89.78 (+1.76)**	91.00	**90.05 (+3.95)**	86.81	**77.93 (+2.31)**	76.62	**82.81 (+6.63)**	70.80
HH, HV, and Location	**89.03 (+1.02)**	90.49	**88.47 (+2.37)**	87.22	**79.78 (+4.16)**	**82.52 (+4.94)**	**81.26 (+5.07)**	**71.36 (+0.76)**
HH, HV, and Time and Location	**91.54 (+3.54)**	**(+88.95) (+2.58)**	**91.65 (+0.91)**	**92.72 (+6.61)**	**78.76 (+3.13)**	75.85	**84.88 (+8.70)**	**72.31 (+1.71)**

could explain why adding AMSR2 89.0 GHz channels might degrade model performance as seen in Table 8.

Compared to the model with only HH and HV as inputs, including all AMSR2 channels improves the SIC, SOD, and FLOE scores by 9.70%, 2.24%, and 3.73% in the testing set as shown in Table 7. Adding a single frequency can achieve similar performance. This is due to the high correlation in brightness-temperature patterns across frequencies, which introduces redundancy in feature space.

Furthermore, the changes in class-wise accuracy after adding or removing AMSR2 channels are investigated for all three parameters. Fig. 8 shows the SOD confusion matrices (in percentage) obtained from the full model (a), the model removing AMSR2 channels (b), and their percentage difference in each element (c). Fig. 8(c) demonstrates that removing AMSR2 impacts most on the classification accuracy of old ice, followed by thick FYI and young ice. Since the other two ice types (new ice and thin FYI) have very few data samples and high ambiguities with neighboring classes, it can be concluded that AMSR2 channels contribute significantly to improving mapping accuracy of major ice types. For SIC, the significant improvement after adding AMSR2 channels is mostly caused by correcting the misclassification of regions with 100% SIC, as shown in Fig. 9. This demonstrates that adding AMSR2 inputs can overcome the ambiguity issue in the electromagnetic signature between open water and sea ice caused by certain ice conditions. An example of this can be seen in Fig. 11. Where HH HV only model misclassifies a large region of 100% ice as open. This scene is in June 23rd, well into the melting season. As for FLOE, the largest improvement caused by adding AMSR2 inputs is observed in the class of vast floe, as shown in Fig. 10. This is also mainly attributed to the reduction of misclassification of ice as open water. Thus, it can be summarized that adding a suitable AMSR2 frequency channel as data inputs improves SOD classification accuracy among most ice types significantly while correcting the misclassification between ice and open water in some regions.

Finally, we aimed to assess how the inclusion or exclusion of passive microwave data influences the model’s seasonal performance. Tables 9 and 10 present the model’s monthly performance metrics, evaluated on the validation dataset, with AMSR2 data either omitted from the full model or incorporated alongside HH and HV polarizations. It is worth noting that Table 9 exclusively reports SOD metrics, as significant variations were observed only in SOD, as elaborated in Table 6. A noticeable decline in model performance for SOD was recorded during March, April, May, and June, when AMSR2 was removed. Highlighting

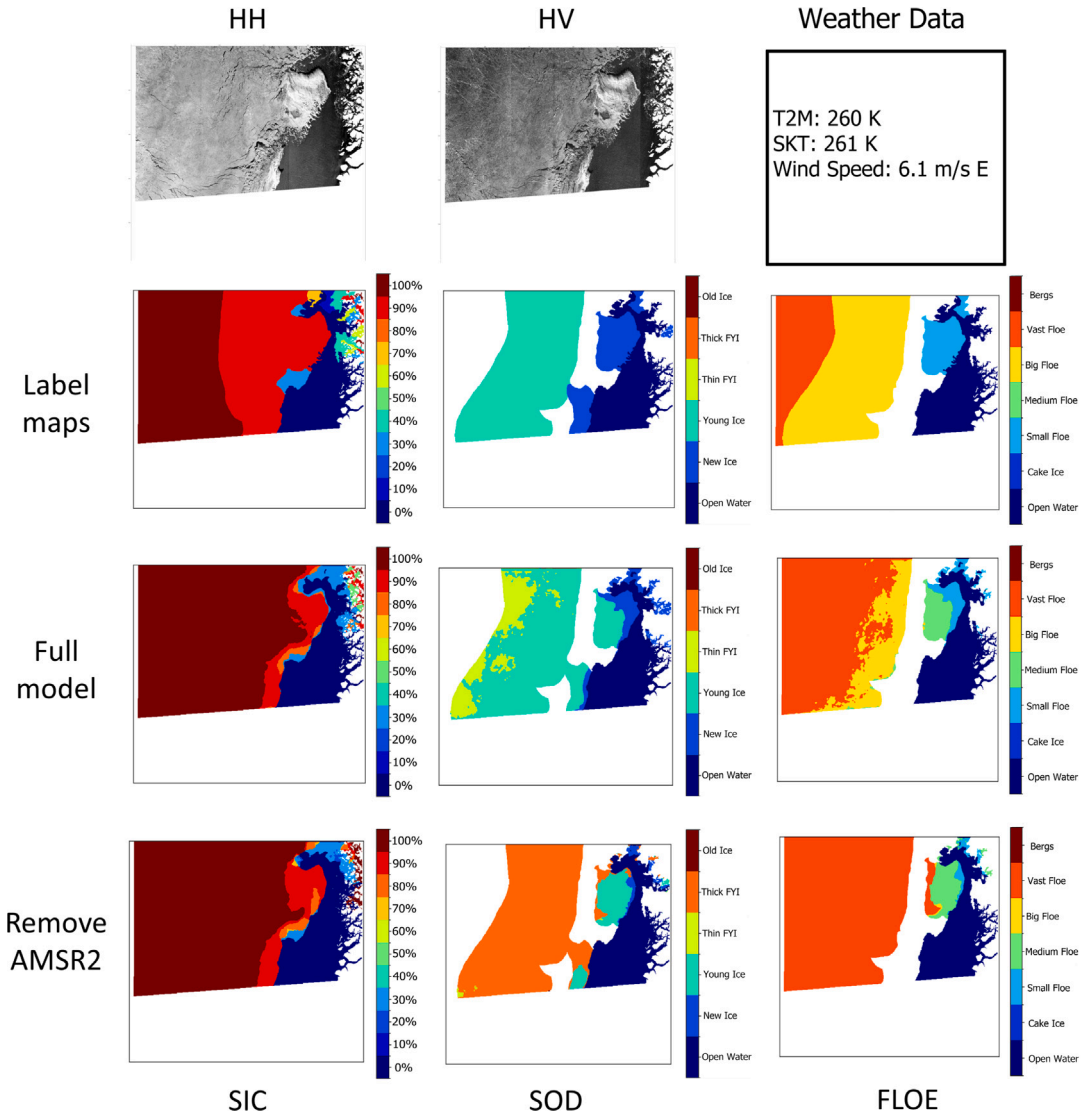


Fig. 4. An example SAR scene in the testing dataset (file ID name: 20211212T211242_dmi, collected on December 12th, 2021 in Northwest Greenland, center latitude/longitude: $71.64°N, 60.53°W$) with corresponding labels (second row) and mapping results (third and fourth rows). The weather parameters like T2M shown in this scene corresponds to the average value across the entire scene. The third row corresponds to results from the full model, while the fourth row corresponds to results from the model without AMSR2 inputs (Remove AMSR2 in Table 6).

that the majority of performance degradation occurs during the melting season. However, it is important to note that no changes in performance were found on the month of July, which is the worst performing month during the melt season for the Full model.

Conversely, the addition of AMSR2 to HH and HV polarizations resulted in an improved SOD performance, particularly in the months of December, January, May, June, and July, as highlighted in Table 10. This improvement is indicative of a more accurate SOD estimation during both the melting and freezing seasons. However, during transitional periods such as October and April, little to no enhancement in SOD performance was observed, pointing to the minimal impact of AMSR2 during these times. For SIC metrics, notable performance improvements were recorded in April, May, and August, emphasizing that the significant benefits of incorporating AMSR2 predominantly occur during the melting season for SIC.

This indicates that AMSR2 significantly contributes to the model's ability to more accurately differentiate between ice thicknesses during the melting periods and freezing periods. On the other hand adding AMSR2 many improves SIC during the melting periods.

Our research findings closely aligns with the existing body of knowledge. Malmgren-Hansen et al. (2021) showed significant improvements in their CNN model for SIC estimation by incorporating AMSR2 data alongside SAR imagery from Sentinel-1. In particular, their enhanced model exhibited a notable increase in accuracy between open water and sea ice by 3.48%, and a 3.62% boost in the accuracy for SIC predictions with the integration of AMSR2 data. The positive impact on SIC estimation is anticipated, considering the prevalent use of AMSR2 in various algorithms to estimating SIC (Comiso and Nishio, 2008).

Moreover, as delineated in Table 8, the inclusion of AMSR2 data enhances the model's performance in estimating the SOD. The literature indicates the potential of passive microwave data in the estimation of the thickness of sea ice. For example, research by Scott et al. (2014) demonstrated that data from AMSR2-E spanning frequencies from 6.9 GHz to 36.5 GHz can predict sea ice thickness up to 50 cm. Moreover, AMSR2 is currently utilized in identifying thin sea ice (≤30 cm) (Cho and Naoki, 2020). This understanding elucidates the observed decline in model performance for classifying new ice (≤10 cm) and young ice (10-30 cm) when AMSR2 data is omitted, as illustrated in Fig. 8.

Table 8
Numerical results obtained from the models (with SAR HH and HV inputs) adding AMSR2 channels individually. The blue and red colors correspond to positive and negative contributions on accuracy, respectively. All results show statistical significance.

Model description	Cross validation (%)				Testing (%)			
	Combined score	SIC	SOD	FLOE	Combined score	SIC	SOD	FLOE
HH, HV only	88.01	90.75	86.10	86.37	75.63	77.58	76.19	70.60
HH, HV, and AMSR2 6.9 GHz	**91.15 (+3.14)**	**92.08 (+1.34)**	**90.13 (+4.03)**	**91.35 (+4.98)**	**81.98 (+6.35)**	**87.68 (+10.10)**	**80.53 (+4.34)**	**73.49 (+2.89)**
HH, HV, and AMSR2 7.3 GHz	**90.59 (+2.57)**	**92.84 (+2.10)**	**89.65 (+3.55)**	**87.95 (+1.58)**	**80.89 (+5.27)**	**87.06 (+9.48)**	**79.27 (+3.09)**	**71.8 (+1.20)**
HH, HV, and AMSR2 10.7 GHz	**90.25 (+2.24)**	**93.02 (+2.27)**	**88.81 (+2.71)**	**87.61 (+1.24)**	**81.07 (+5.44)**	**87.12 (+9.54)**	**79.43 (+3.24)**	**72.24 (+1.65)**
HH, HV, and AMSR2 18.7 GHz	**90.15 (+2.13)**	**93.2 (+2.46)**	**88.25 (+2.14)**	**87.84 (+1.47)**	**80.88 (+5.25)**	**87.15 (+9.57)**	**78.61 (+2.42)**	**72.89 (+2.29)**
HH, HV, and AMSR2 23.8 GHz	**90.18 (+2.17)**	**92.8 (+2.05)**	**88.53 (+2.43)**	**88.26 (+1.89)**	**80.44 (+4.81)**	**86.92 (+9.34)**	**77.48 (+1.29)**	**73.39 (+2.79)**
HH, HV, and AMSR2 36.5 GHz	**90.3 (+2.29)**	**92.56 (+1.81)**	**89.06 (+2.96)**	**88.27 (+1.90)**	**81.01 (+5.38)**	**87.84 (+10.27)**	**77.68 (+1.49)**	**73.99 (+3.39)**
HH, HV, and AMSR2 89.0 GHz	**84.51 (−3.50)**	**88.79 (−1.96)**	**81.65 (−4.45)**	**81.66 (−4.71)**	**80.15 (+4.52)**	**85.83 (+8.26)**	**77.57 (+1.39)**	**73.94 (+3.34)**

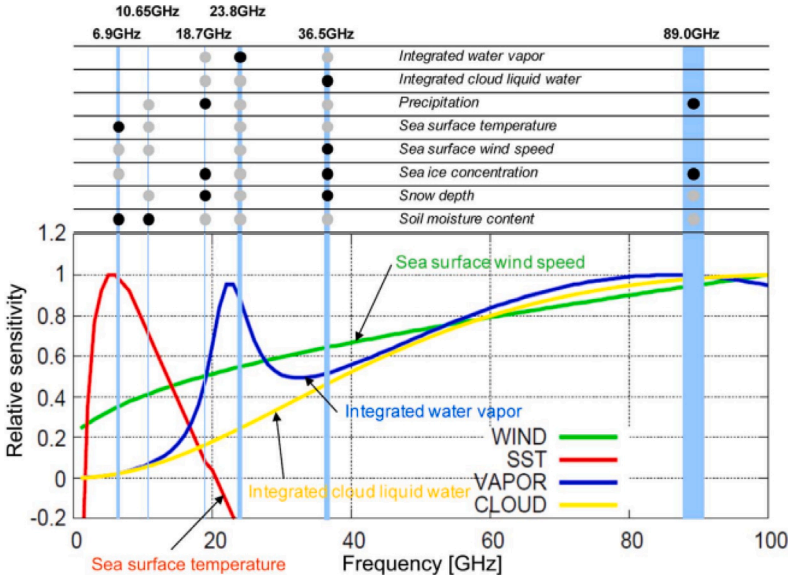


Fig. 5. Relative sensitivity of AMSR2 for oceanic geophysical parameters as a function of observing frequency (Imaoka et al., 2010).

Table 9
Comparison of the monthly performance on SOD between "Full Model" and "Remove AMSR2". Statistical significant decreases are shown in bold.

Months	Full model			Remove AMSR2			Decrease
	SOD	Std	Sample number	SOD	Std	Sample number	
January	82.98	18.34	45	80.96	17.32	52	−2.02
February	86.82	17.46	44	85.15	19.74	47	−1.68
March	88.53	11.92	32	**81.94**	21.15	45	**−6.59**
April	91.60	9.33	47	**86.89**	12.72	72	**−4.71**
May	94.65	7.92	44	**87.27**	14.45	53	**−7.38**
June	93.84	6.21	37	**88.82**	9.54	58	**−5.03**
July	86.95	13.36	61	87.12	18.48	67	0.18
August	91.25	11.27	97	90.18	13.39	92	−1.07
September	93.44	8.62	82	91.74	10.47	107	−1.70
October	89.76	14.94	67	92.22	10.12	77	2.46
November	88.30	15.22	45	86.43	19.83	63	−1.87
December	79.11	22.76	53	80.62	23.58	63	1.51

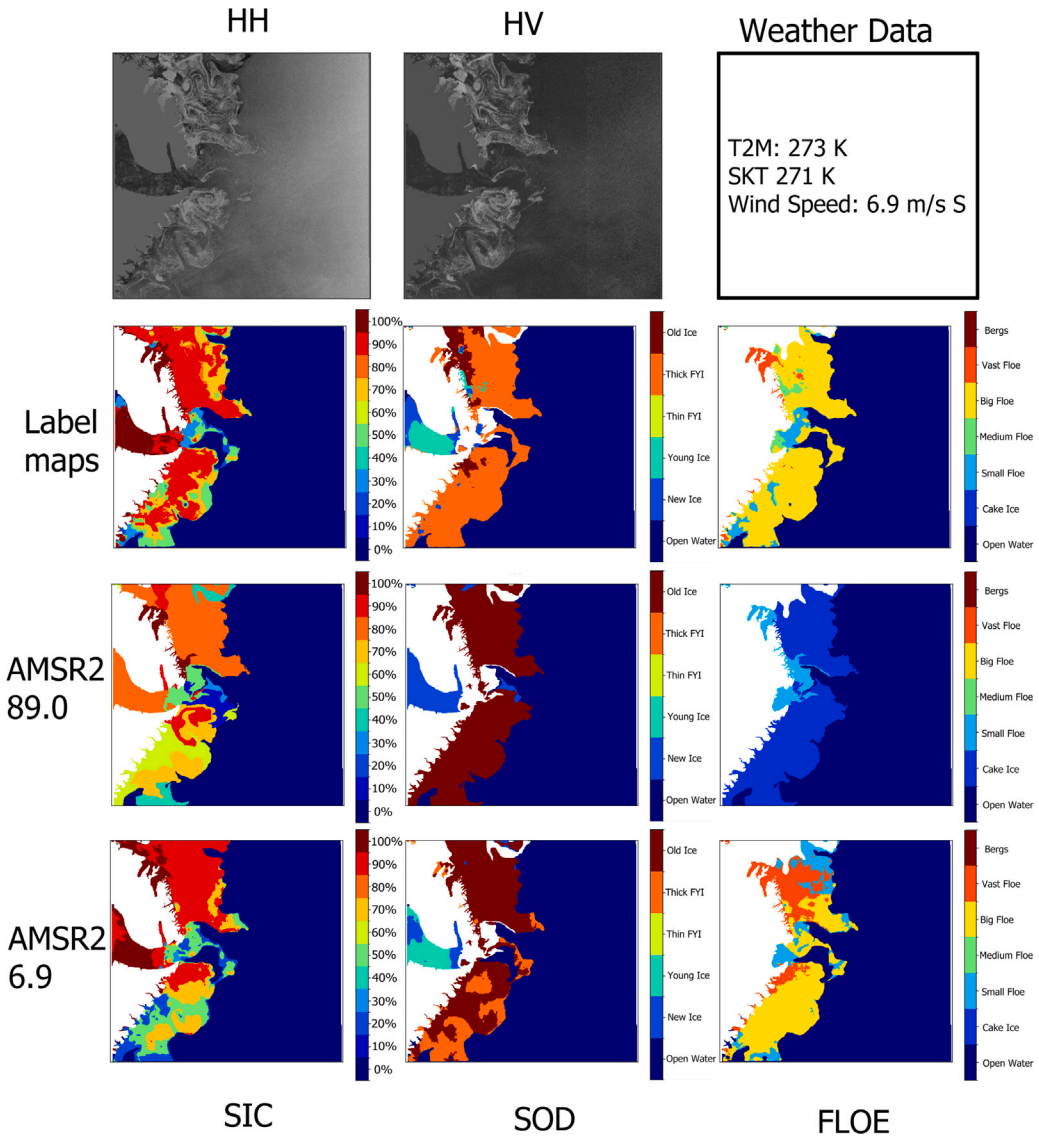


Fig. 6. An example comparing the mapping results from models with two different feature combinations: HH + HV + AMSR2 89.0 GHz and HH + HV + AMSR2 6.9 GHz. File ID name of the SAR scene: 20201013T080448_dmi, collected on October 10th, 2020 in the east coast of Greenland, center latitude/longitude: $70.10°N, 19.74°W$. The weather parameters like T2M shown in this scene corresponds to the average value across the entire scene.

Table 10
Comparison of the monthly performance on SIC and SOD between "HH HV Only" and "HH HV and AMSR2". Statistical significant improvements are shown in bold.

Months	HH HV Only		HH HV and AMSR2		Increase	
	SIC	SOD	SIC	SOD	SIC	SOD
January	91.13	67.15	93.00	**80.84**	+1.88	+**13.69**
February	97.01	79.59	96.64	83.95	−0.37	+4.37
March	95.03	77.09	94.76	84.04	−0.27	+6.96
April	89.78	87.58	**93.72**	86.40	+**3.95**	−1.18
May	90.54	85.50	**92.56**	**90.26**	+**2.02**	+**4.76**
June	88.18	87.18	91.07	**94.18**	+2.89	+**7.00**
July	84.39	83.40	89.18	**89.43**	+4.78	+**6.02**
August	90.44	86.22	**93.08**	89.27	+**2.64**	+3.05
September	94.60	92.00	94.95	93.74	+0.35	+1.74
October	91.76	88.44	92.95	88.19	+1.19	−0.25
November	92.04	85.54	94.33	86.63	+2.29	1.09
December	93.85	69.15	95.28	**77.27**	+1.44	+**8.12**

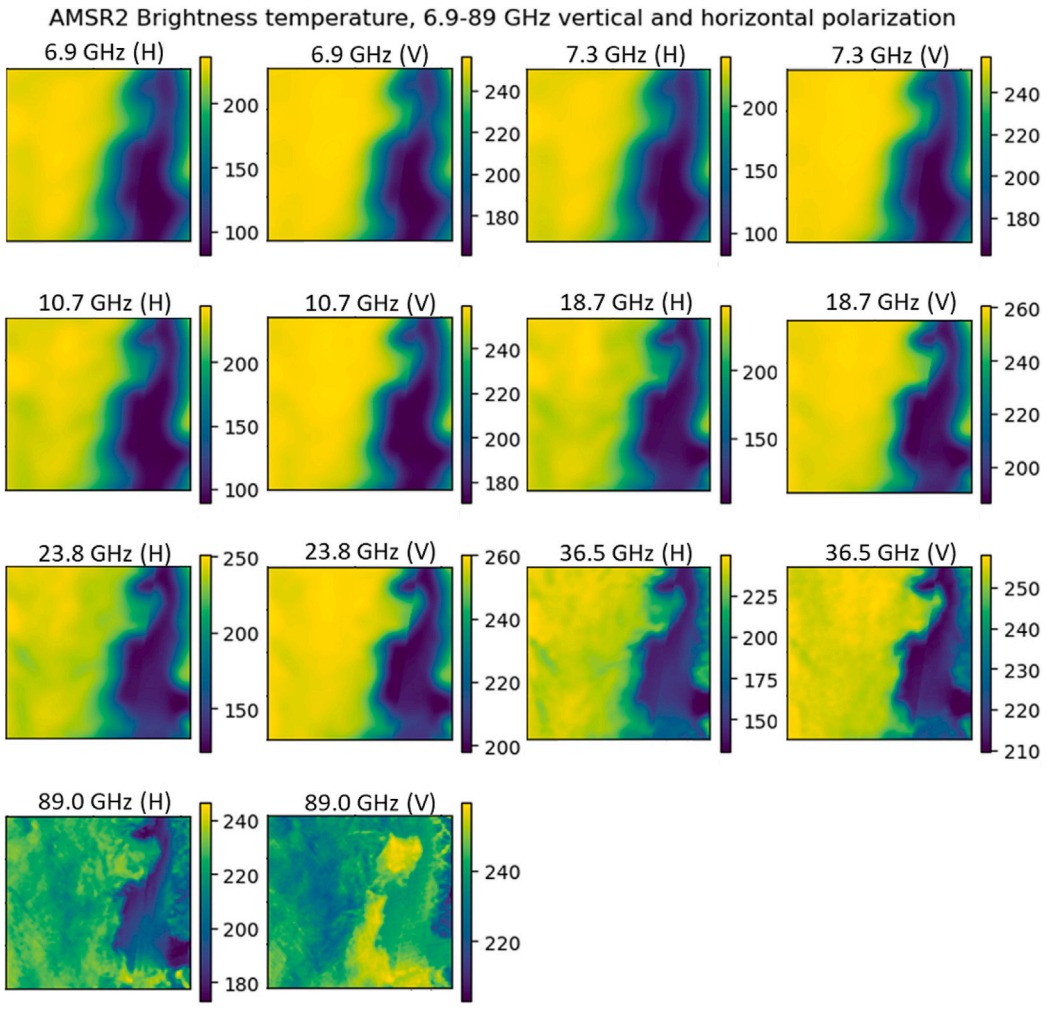


Fig. 7. The corresponding AMSR2 feature maps of the SAR scene in Fig. 4 in all available frequencies.

Table 11
Comparison of the monthly performance on SOD between "HH HV Only" and "HH HV + Time". Statistical significant improvements are shown in bold.

Months	HH HV only			HH HV + time			Improvement
	SOD	Std	Sample number	SOD	Std	Sample number	
January	67.15	28.15	64	**85.12**	14.68	47	**+17.98**
February	79.59	24.83	49	82.12	22.22	43	+2.53
March	77.09	27.24	46	**87.44**	12.46	50	**+10.36**
April	87.59	11.41	61	**90.93**	8.90	40	**+3.35**
May	85.50	16.64	73	**92.66**	8.85	44	**+7.15**
June	87.18	9.30	43	87.67	16.95	35	+0.49
July	83.40	17.65	65	83.24	21.05	57	−0.16
August	86.22	18.72	99	88.73	17.60	93	+2.51
September	92.00	13.66	113	93.64	8.34	85	+1.64
October	88.44	17.30	80	**92.82**	9.99	71	**+4.37**
November	85.54	19.74	66	87.42	18.93	57	+1.88
December	69.15	29.15	61	**83.03**	16.39	38	**+13.88**

4.4. Analysis on spatial–temporal features

Examples of comparison between mapping results obtained from the full model and the model removing location and/or time inputs are depicted in Figs. 12 and 13, which all demonstrate the effectiveness of those features in improving SOD classification accuracy. For instance, in Fig. 13, the full model is able to detect the presence of old ice, while the model without location and time information inputs totally misclassifies those pixels.

Furthermore, Fig. 14 shows the difference in each element of the SOD class-wise confusion matrix between the full model and the model removing time and/or location information. Compared to the full model, removing time inputs reduces the classification accuracy of new ice (−13%), young ice (−18%), and old ice (−13%) significantly (Fig. 14(a)). This can be explained by the fact that new and young are mostly present in the freeze-up season and adding time inputs facilitates the model to learn such patterns. Table 11 supports this hypothesis by comparing the monthly SOD score between HH HV only and HH HV + Time. The bold text indicates statistical significance. The performance of the model drastically increases in December and January when time is introduced, indicating that much of the performance is gained during the freeze-up season.

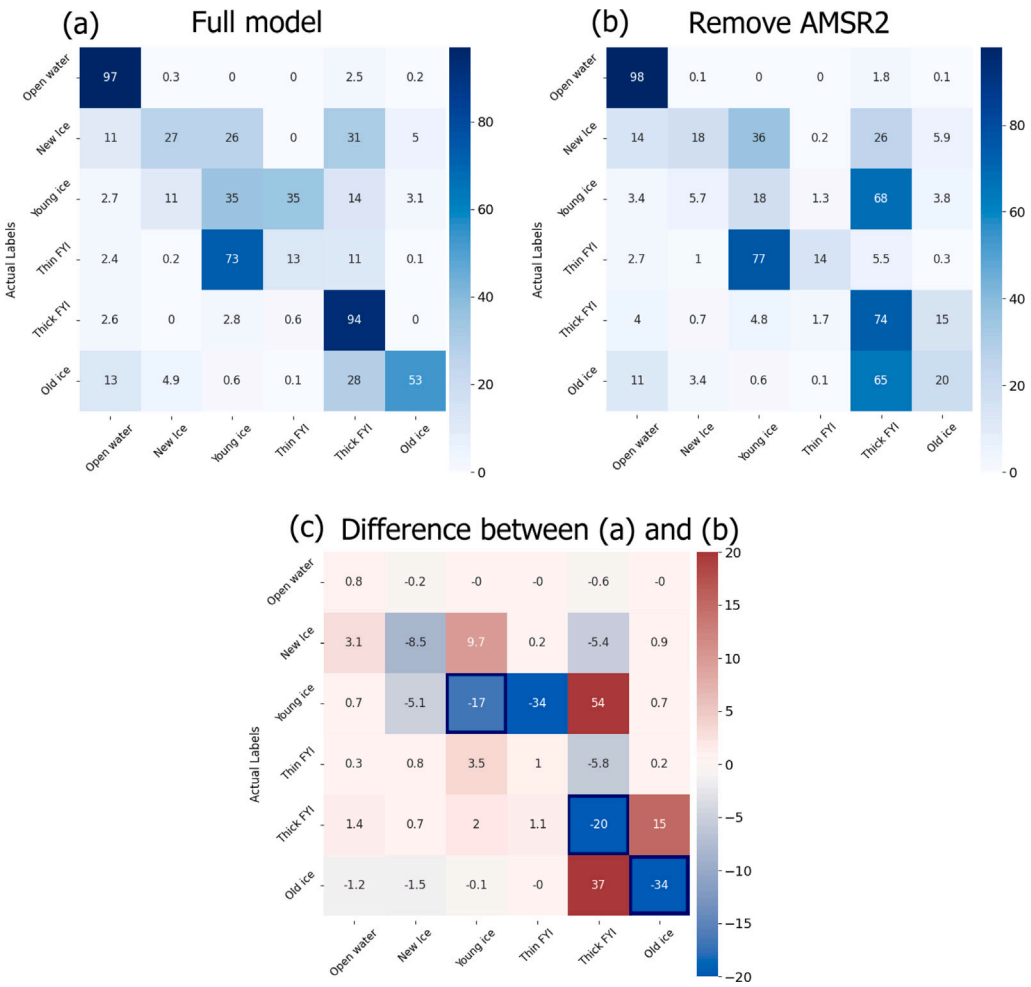


Fig. 8. (a) The confusion matrix of SOD classification results (testing set) from the full model. (b) The confusion matrix of SOD classification results (testing set) from the model removing AMSR2 features ("Remove AMSR2" in Table 6). (c) The value difference in each element between (b) and (a).

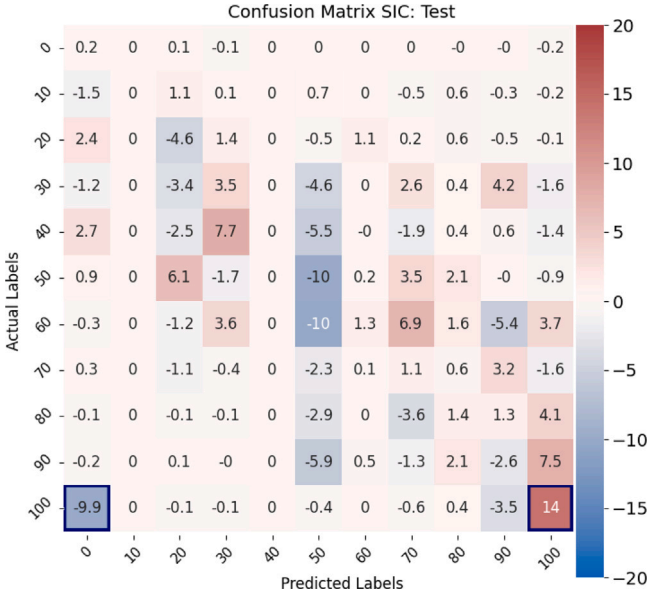


Fig. 9. The value change in each element of the SIC classification confusion matrix test dataset after adding AMSR2 features, in comparison with the "HH, HV only" model.

In contrast, removing location information inputs mainly reduces the classification accuracy of thick FYI (−13%) and old ice (−17%). The possible reason is the relatively stable spatial distributions of sea ice extent and ice types. For example, old (multi-year) ice mainly occurs in the Canadian archipelago and east coast of Greenland. In consequence, adding location information inputs enables the model to learn about the spatial distributions of different ice types. Removing both time and spatial information further reduces the classification accuracy of most ice types, as demonstrated in Fig. 14(c).

4.5. ERA5 feature analysis

To investigate the effect of different ERA5 features on model performance, each individual feature is added to the "HH, HV only" model for comparison, as presented in Table 12. A slight improvement in cross-validation accuracy can be observed in all models with ERA5 feature inputs. However, only the model that incorporates wind speed input improves the accuracy of the test, while others show lower scores, especially in SIC. Therefore, we look into each SAR scene and find the one with low SIC scores, as depicted in Fig. 15. The corresponding ERA5 feature maps of this SAR scene is presented in Fig. 16. Fig. 15(d) shows that a large area of ice regions with 100% SIC are misclassified as open water from "HH, HV only" model, which can be caused by the melting on ice surface. In consequence, the patterns of ice in SAR imagery become blurry. Adding non-wind ERA5 features (e.g., T2M in Fig. 15(e)) can deteriorate this issue with even larger areas being misclassified. This issue is alleviated significantly after adding wind

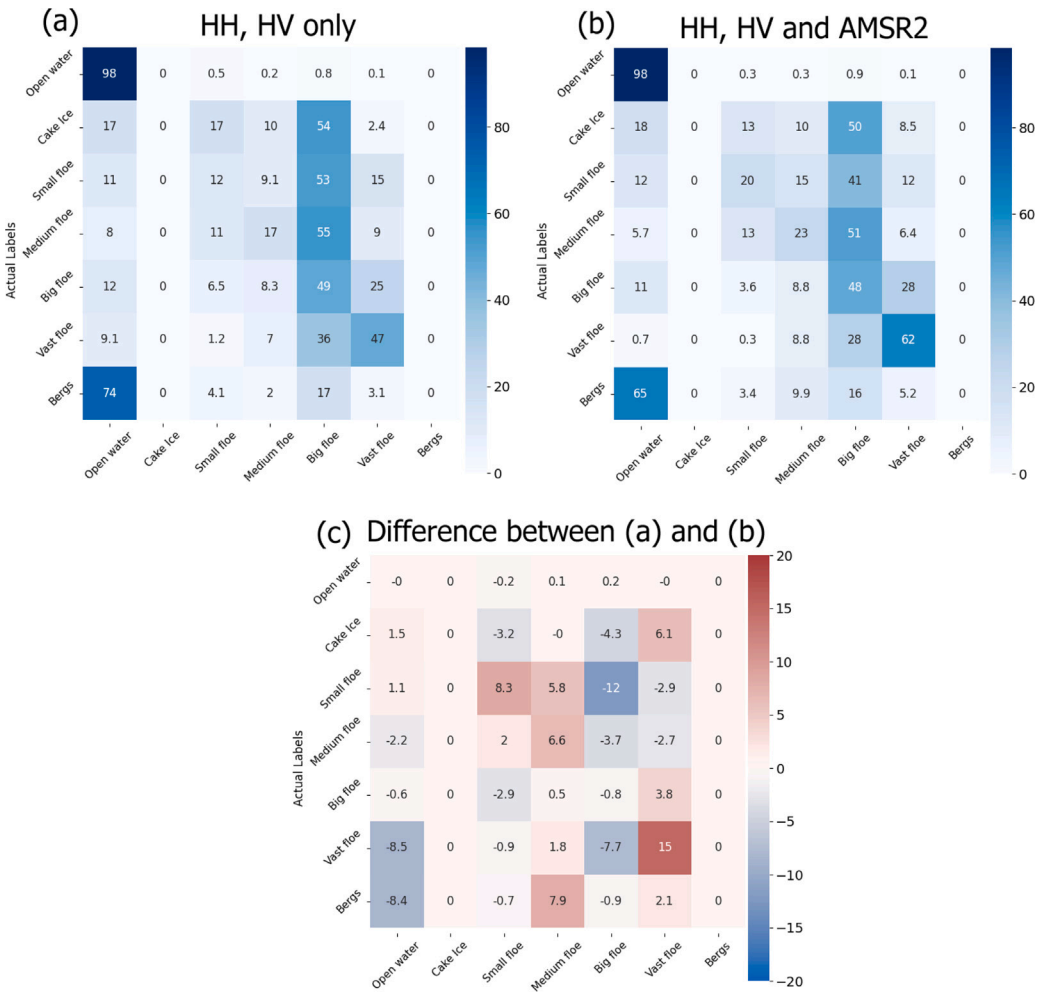


Fig. 10. (a) The confusion matrix of FLOE classification results (testing set) from the "HH, HV only" model. (b) The confusion matrix of FLOE classification results (testing set) from the model adding AMSR2 features ("HH, HV, and AMSR2" in Table 7). (c) The value difference in each element between (b) and (a).

Table 12
Numerical results obtained from the models (with SAR HH and HV inputs) adding ERA5 features individually. The blue and red colors correspond to positive and negative contributions on performance, respectively.

Model description	Cross validation (%)				Testing (%)			
	Combined score	SIC	SOD	FLOE	Combined score	SIC	SOD	FLOE
HH, HV Only	88.01	90.74	86.10	86.37	75.63	77.57	76.19	70.60
HH, HV, and wind speed	88.56	90.92	86.88	87.22	**79.94 (+4.31)**	**86.71 (+9.13)**	**77.15 (+0.96)**	**71.97 (+1.38)**
HH, HV, and SKT	**88.90 (+0.89)**	**91.61 (+0.87)**	**87.13 (+1.03)**	87.04	**71.43 (−4.20)**	**67.82 (−9.76)**	76.35	**68.81 (−1.79)**
HH, HV, and T2M	88.65	91.11	**87.38 (+1.27)**	86.26	**71.96 (−3.67)**	**69.13 (−8.44)**	76.16	**69.21 (−1.38)**
HH, HV and TCWV, TCLW	88.45	91.12	86.75	86.49	**71.62 (−4.00)**	**69.70 (−7.87)**	**75.05 (−1.14)**	**68.60 (−2.00)**

speed as data inputs, as manifested in Figs. 15(f) and 17 (classification accuracy of 100% SIC increase by 16%). Thus, it can be concluded that incorporating wind information as data inputs could further improve model performance in SIC estimation under challenging circumstances such as melting on ice surface, while other weather parameters might not be helpful in this case.

Furthermore, as shown in Table 7 adding all ERA5 features significantly improves the testing performance. On the contrary, removing the ERA5 features from the full model as shown in Table 6 does not significantly reduce the scores in either testing or cross validation. Indicating that the ERA5 features are redundant when the AMSR2 features are already present. These findings contradict what is found

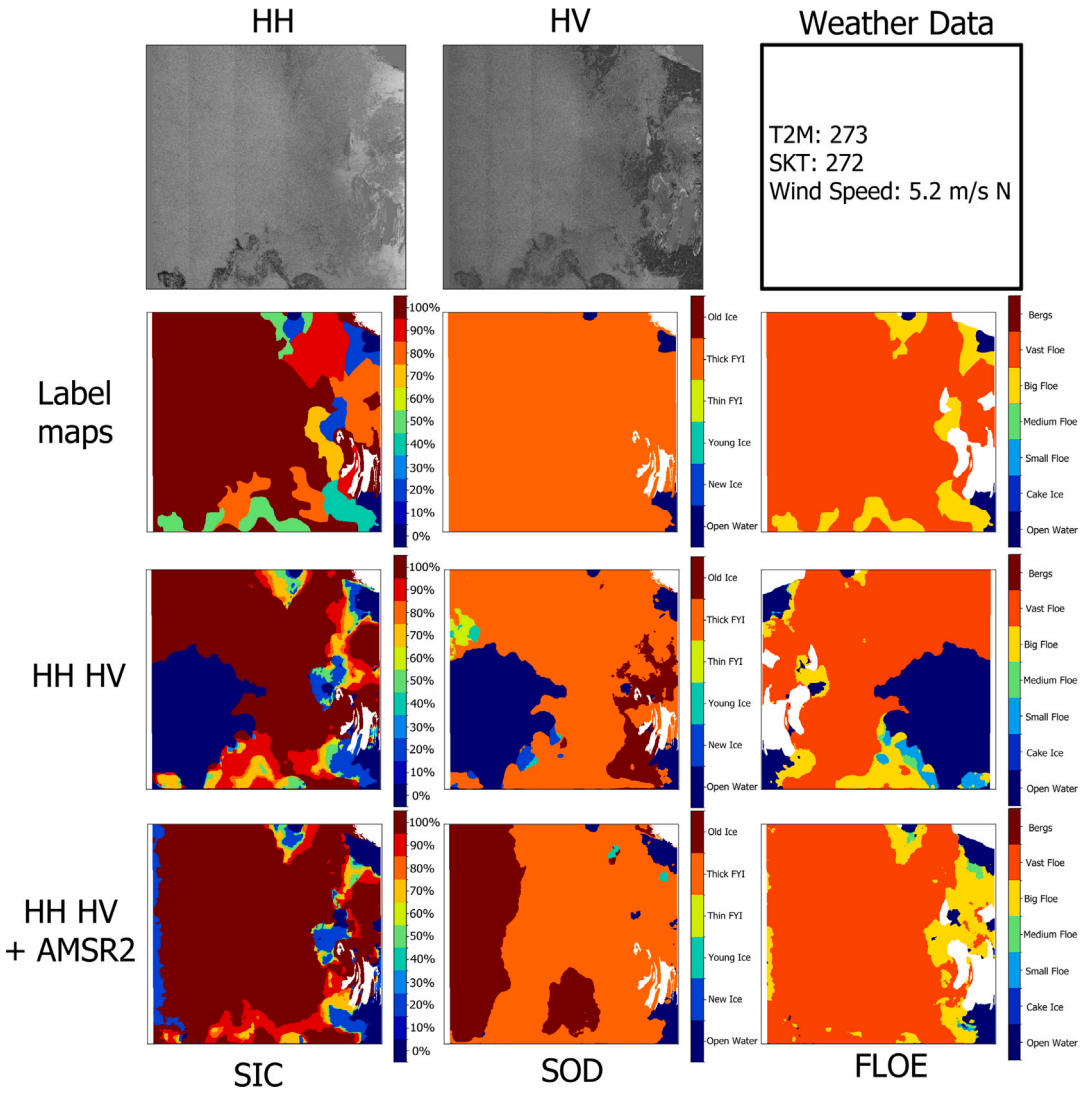


Fig. 11. An example SAR scene in the testing dataset under melting conditions(file ID name: 20180623T114935_cis, collected on June 23rd, 2018 in Hudson Bay, near the Belcher Islands, center latitude/longitude: $57.32°N, 81.98°W$) with corresponding labels (second row) and mapping results (third and fourth rows). The weather parameters like T2M shown in this scene corresponds to the average value across the entire scene. The third row corresponds to results from the HH HV only model, while the fourth row corresponds to results from HH HV and AMSR2 model (HH, HV, and AMSR2 in Table 7).

in the literature. Algorithms that utilize AMSR2 normally benefit from the addition of ERA5 weather data. Han et al. (2021a) used a random forest algorithm with AMSR2 data + ERA5 data to improve the performance of SIC estimation. They found that their algorithm using AMSR2 + ERA5 had at least an 8.3% increase in R^2 score over other algorithms that only utilized AMSR2 data. Furthermore, Nihashi et al. found that adding ERA5 data in conjunction to AMSR2 improved the thickness estimation of thin sea ice (Nihashi et al., 2024). However, it is important to note that none of these studies used SAR imagery data. It can be hypothesized that when both SAR imagery and AMSR2 data are present the addition of ERA 5 weather data becomes unnecessary.

4.6. Feature importance ranking

The ranking for the relative importance of different data inputs could provide valuable guidance for operational sea ice mapping. Thus, based on the detailed analysis above, we implement a relative importance score (RIS) to evaluate each feature. For example, the RIS of AMSR2 6.9 GHz in SIC (RIS_{SIC}) can be expressed as

$$RIS_{SIC} = SIC^{val}_{HH,HV,6.9GHz} - SIC^{val}_{HH,HV} + SIC^{test}_{HH,HV,6.9GHz} - SIC^{test}_{HH,HV}, \quad (7)$$

where $SIC^{val}_{HH,HV,6.9GHz}$ and $SIC^{test}_{HH,HV,6.9GHz}$ correspond to the SIC scores of the model with SAR HH, HV, and AMSR2 6.9 GHz inputs in cross validation and testing sets. $SIC^{val}_{HH,HV}$ and $SIC^{test}_{HH,HV}$ refer to the scores of the model with SAR HH and HV inputs only. All the scores used for calculation can be found in Table 7. This equation is also implemented to calculate the RIS of the combined scores, SOD, and FLOE. The equation indicates that the larger the improvement compared to the "HH, HV only" model, the more important the feature is.

The rankings calculated from Eq. (7) are summarized in Table 13. Features that improve the scores significantly (here determined as RIS values higher than 3%) are highlighted in blue. In contrast, features that reduce the scores (here determined as negative RIS values) are highlighted in red. Features with positive RIS but less than 3% are not

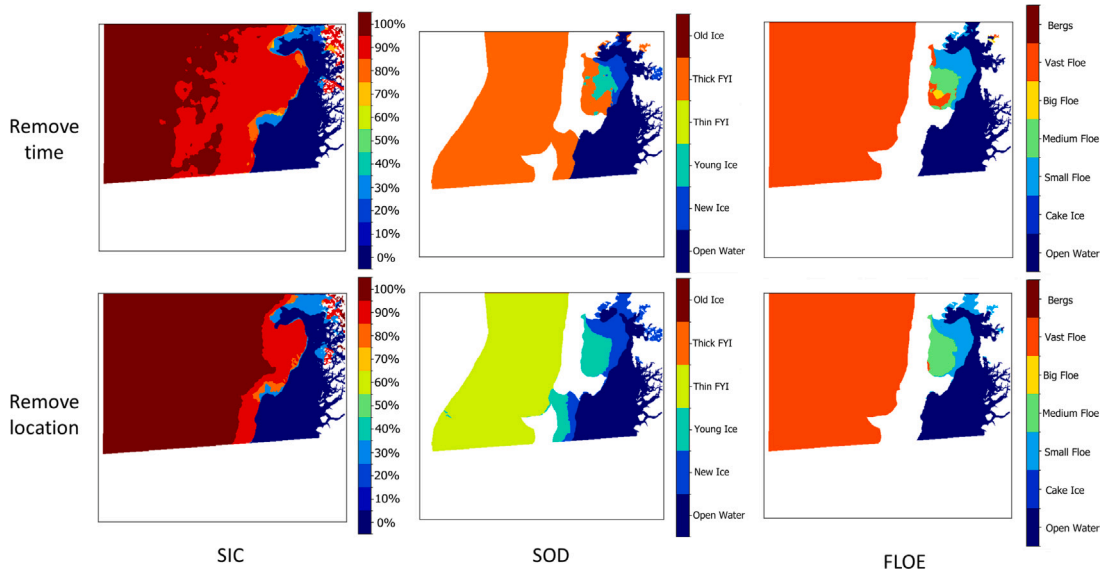


Fig. 12. Mapping results of the SAR scenes in Fig. 4 from more models. First row: the mapping results from the model without time information inputs ("Remove Time" in Table 6). Second row: the mapping results from the model without location information inputs ("Remove Location" in Table 6).

Table 13
A ranking of importance for different features in the combined score and individual parameters. Features that show significant improvement, no significant effects, and negative effects in mapping accuracy are in blue, black, and red, respectively.

Combined score ranking	SIC score ranking	SOD score ranking	FLOE score ranking
AMSR2 6.9 GHz	AMSR2 36.5 GHz	Time information	AMSR2 6.9 GHz
AMSR2 7.3 GHz	AMSR2 18.7 GHz	AMSR2 6.9 GHz	AMSR2 36.5 GHz
AMSR2 10.7 GHz	AMSR2 10.7 GHz	Location information	AMSR2 23.8 GHz
AMSR2 36.5 GHz	AMSR2 7.3 GHz	AMSR2 7.3 GHz	AMSR2 18.7 GHz
AMSR2 18.7 GHz	AMSR2 6.9 GHz	AMSR2 10.7 GHz	AMSR2 10.7 GHz
AMSR2 23.8 GHz	AMSR2 23.8 GHz	AMSR2 18.7 GHz	AMSR2 7.3 GHz
Location information	Wind	AMSR2 36.5 GHz	Wind
Wind	AMSR2 89.0 GHz	AMSR2 23.8 GHz	Distance
Time information	Location information	Wind	Location information
AMSR2 89.0 GHz	IA	Distance	Time information
IA	Distance	T2M	IA
Distance	Time information	IA	Skin Temperature
T2M	T2M	Skin Temperature	AMSR2 89.0 GHz
Skin Temperature	Skin Temperature	TCWV, TCLW	T2M
TCWV, TCLW	TCWV, TCLW	AMSR2 89.0 GHz	TCWV, TCLW

highlighted. It can be observed that the rankings are consistent with our findings above. AMSR2 features (except 89.0 GHz) play a vital role in improving the accuracy of all parameters. Time and location information are the most effective data inputs in boosting SOD classification accuracy. Wind data could effectively improve SIC estimation accuracy. Features including distance-to-land map and incidence angle do not seem to affect model performance, while others such as AMSR2 89.0 GHz and non-wind ERA5 data could negatively affect model accuracy.

It is important to note that while ASMR2 channels highly improve the model's performance, adding more than one AMSR2 frequency may not improve the performance of the model due to redundant data. As such it is advised to add instead data sources from different types. Here our recommendation would be to utilize one frequency channel from AMSR2 and time and location channels.

5. Conclusion and future work

To provide insights and guidance for automated sea ice mapping from SAR imagery and multi-source data, this paper conducts a comparative study to evaluate the importance of various data inputs on sea ice parameter retrieval. The AI4Arctic dataset with SAR imagery and corresponding AMSR2, ERA5, ancillary data is utilized to train multitask models based on a U-Net architecture. For feature importance analysis, models with different combinations of data inputs are compared with each other via ablation studies and the alternate adding of individual data input.

The ablation studies demonstrate that compared to the model incorporating all available inputs, removing AMSR2, time and location information channels has a significant impact on model performance, especially the classification accuracy of major ice types (young ice, thick FYI, and old ice) in SOD. Besides, time information further boosts

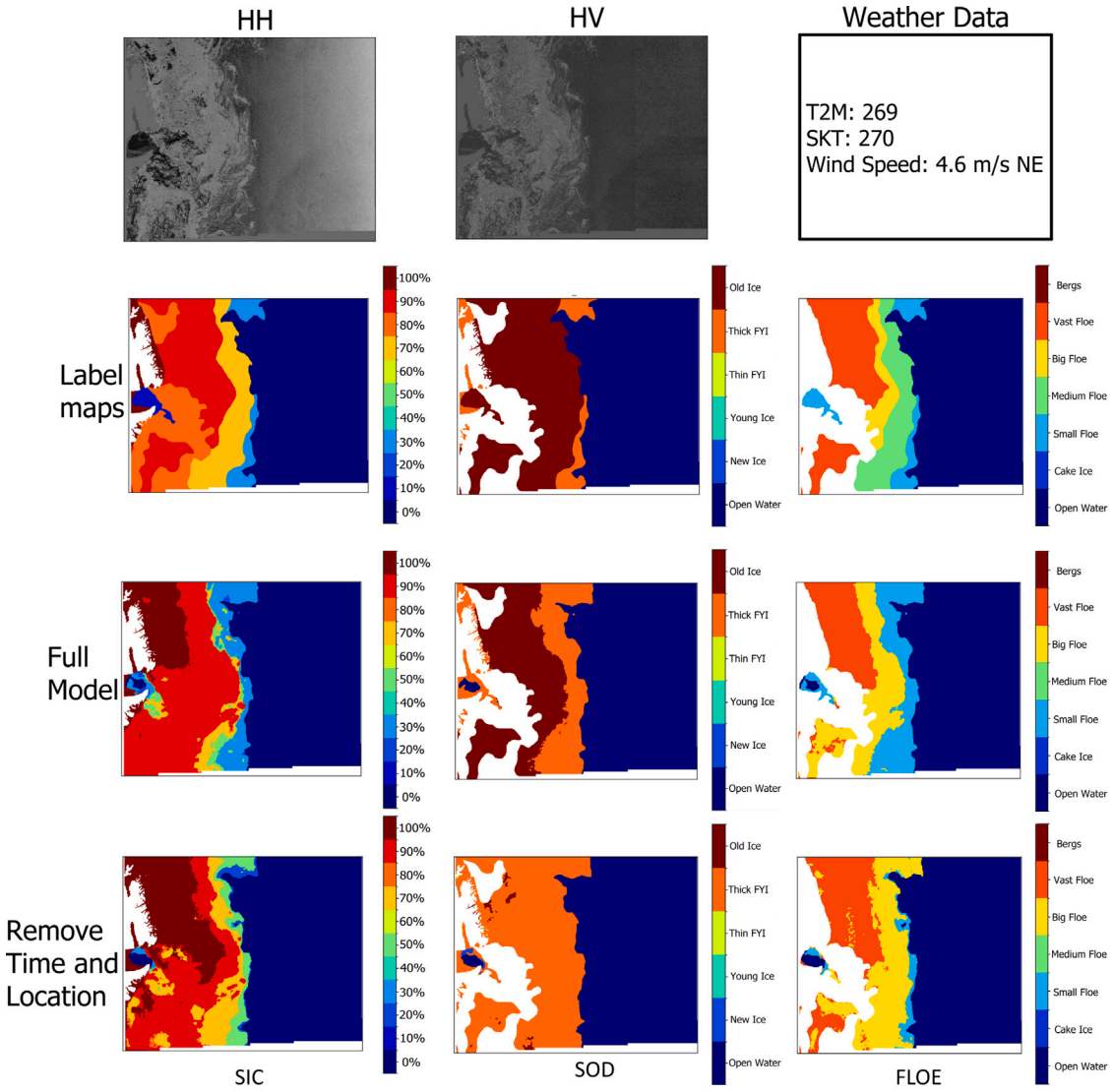


Fig. 13. An example SAR scene in the testing dataset (file ID name: 20210506T075557_dmi, collected on May 6th, 2021 in Central East Greenland, center latitude/longitude: $68.89°N, 17.49°W$) with corresponding labels (second row) and mapping results (third and fourth rows). The weather parameters like T2M shown in this scene corresponds to the average value across the entire scene. The third row corresponds to results from the full model, while the fourth row corresponds to results from the model without time and location inputs ("Remove time and location" in Table 6).

the classification of ice types that are time-sensitive (e.g., new and young ice), while location information facilitates the model to have a better knowledge of the spatial distributions of FYI and old ice. The removal of other feature types (e.g., ERA5 and IA) does not reduce accuracy scores, which indicates their redundancy among all available features.

On the other hand, we also evaluate the impact of adding individual feature on the model with dual-polarized SAR imagery as inputs. Results validate the effectiveness of most of the AMSR2 channels (except 89.0 GHz) on improving the mapping accuracy of all the three sea ice parameters (SIC, SOD, and FLOE). Nevertheless, adding multiple AMSR2 channels with different frequencies might not further improve model performance due to information redundancy. Among all the ERA5 weather parameters, adding wind speed data inputs improves SIC estimation accuracy by correcting the misclassification of ice regions, whereas other features might negatively affect estimation accuracy. Finally, based on the statistical results, a metric is proposed to rank the relative importance of all auxiliary data inputs in sea ice mapping, which helps us identify features that are useful in enhancing model performance.

Despite the comprehensive comparison and analysis, the findings should be further validated by more data and more advanced DL-based models in future works. For example, the model should be trained and evaluated again once the next version of the AI4Arctic dataset is released. The new dataset will comprise 16 times as much data compared to the competition dataset with expanded the geographical coverage (Stokholm et al., 2023). Ice charts are produced by manual interpretation of SAR imagery and various data sources. Although they are used as ground truths here for model training, they only provide regional information of sea ice parameters with different sources of uncertainties, including limitations in spatial and temporal resolution, operator biases, and errors in setting ice concentration labels for intermediate ice concentrations. Thus, reference data from other sources should be incorporated. In particular, edge specific metrics such as the one proposed in Melsom et al. (2019) should be introduced to evaluate model performance around ice-water boundaries once detailed labels

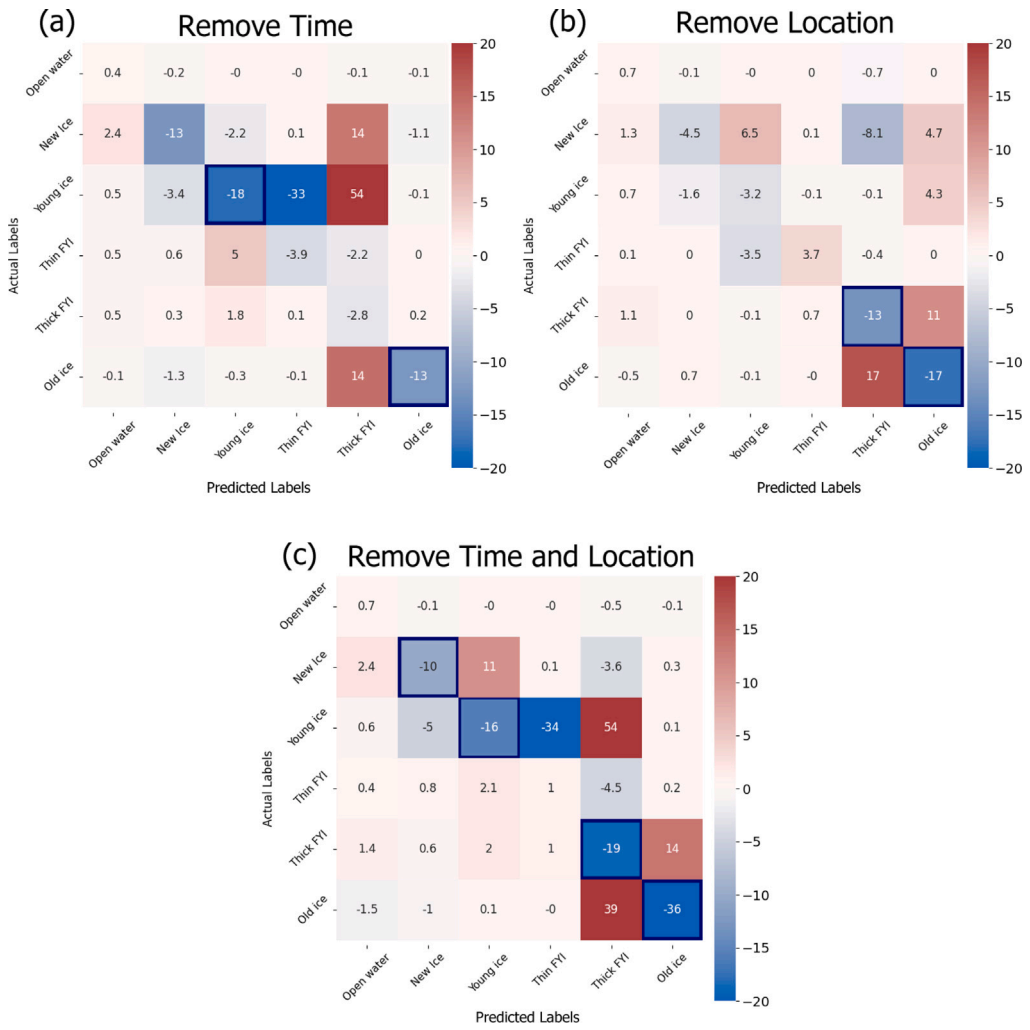


Fig. 14. (a) The value change in each element of the SOD classification confusion matrix after removing time inputs from the full model. (b) The value change in each element of the SOD classification confusion matrix after removing location inputs from the full model. (c) The value change in each element of the SOD classification confusion matrix after removing time and location inputs from the full model.

with high resolution are available. Compared to other ice types, new ice and thin FYI consists of very small portions from the whole data and this could be one reason why the current model struggles to identify them correctly. Thus, it would be worthwhile to further validate this once data collected in regions with significant new ice and new FYI formation (e.g., the Barents and Kara seas) are available. In addition, data collected from other satellite sensors (e.g., optical, SAR images from other bands such as L-band) should be incorporated for analysis to see if they will further improve classification accuracy, and particularly for tactical navigation applications, without the penalty of downgraded resolution and edge fidelity incurred by the incorporation of AMSR2. Finally, an evaluation of the performance of the model across different regions of the Arctic should be done to understand which areas are challenging for the model.

CRediT authorship contribution statement

Xinwei Chen: Writing – review & editing, Writing – original draft, Visualization, Supervision, Methodology, Investigation, Formal analysis. **Fernando J. Pena Cantu:** Writing – review & editing, Visualization, Methodology, Investigation, Formal analysis, Conceptualization. **Muhammed Patel:** Writing – review & editing, Methodology, Conceptualization. **Linlin Xu:** Writing – review & editing, Supervision. **Neil C. Brubacher:** Writing – review & editing. **K. Andrea Scott:** Writing – review & editing, Supervision. **David A. Clausi:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data used in this research is the AI4Arctic Sea Ice Challenge Dataset provided by Technical University of Denmark (data link: https://data.dtu.dk/collections/AI4Arctic_Sea_Ice_Challenge_Dataset/6244065/2). The code of this research is available at https://github.com/echonax07/MMSeaIce.

Acknowledgments

The authors acknowledge the providers of the AI4Arctic dataset and the organizers of the AutoIce competition. This work was supported in part by Environment and Climate Change Canada (ECCC) and the Natural Sciences and Engineering Research Council (NSERC) of Canada. This research was enabled in part by support provided by Calcul Québec's (calculquebec.ca) and the Digital Research Alliance of Canada (alliancecan.ca).

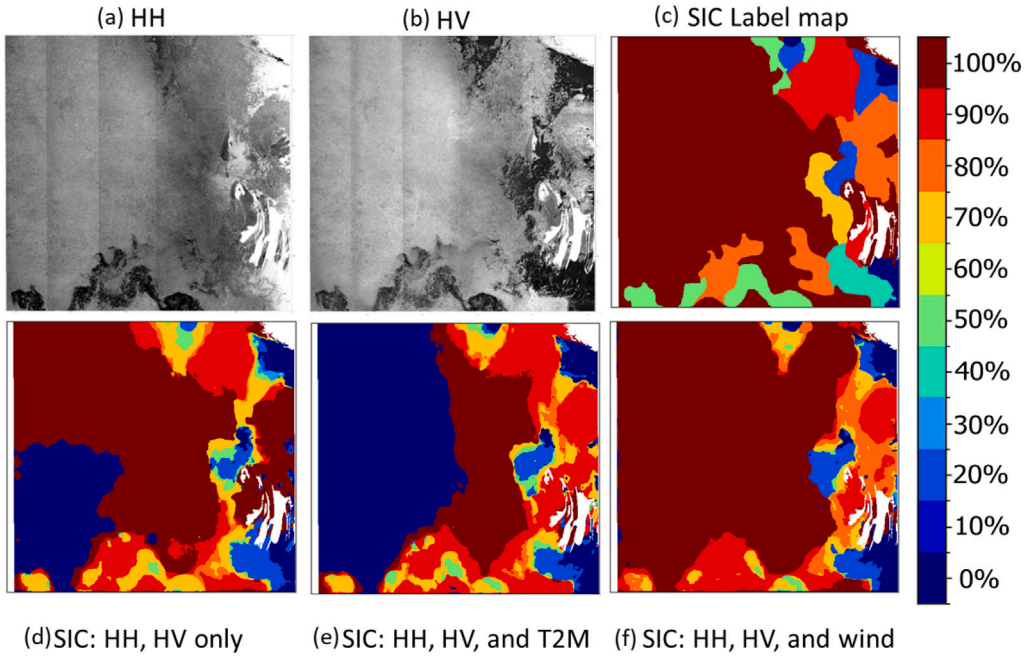


Fig. 15. An example SAR scene in the testing dataset (file ID name: 20180623T114935_cis, collected on June 23rd, 2018 in Hudson Bay, near the Belcher Islands, center latitude/longitude: $57.32°N, 81.98°W$), with HH and HV channels depicted in (a) and (b), respectively. (c) The corresponding SIC label map. (d) Mapping results from "HH, HV only" model. (e) Mapping results from "HH, HV, and T2M" model in Table 12. (f) Mapping results from "HH, HV, and wind speed" model in Table 12.

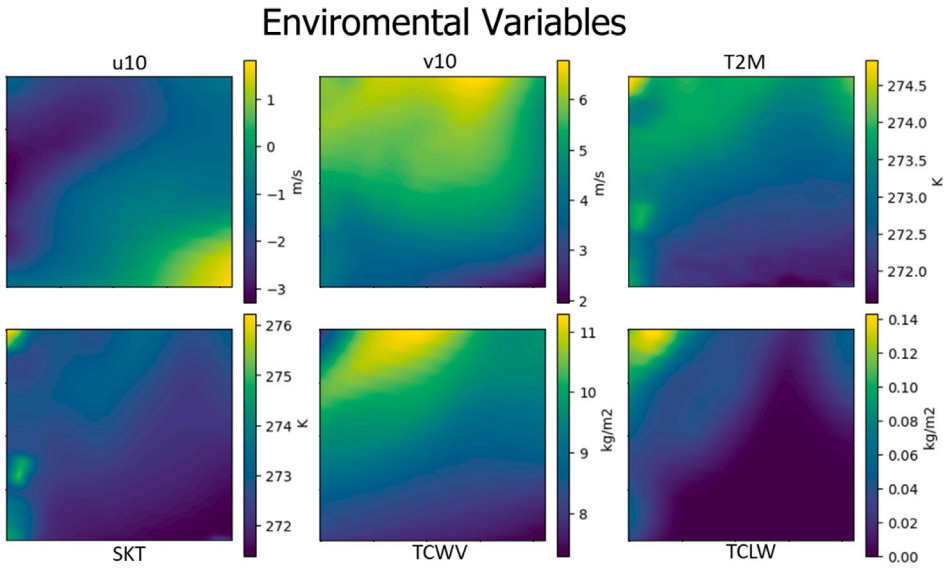


Fig. 16. The corresponding ERA5 feature maps of the SAR scene in Fig. 15.

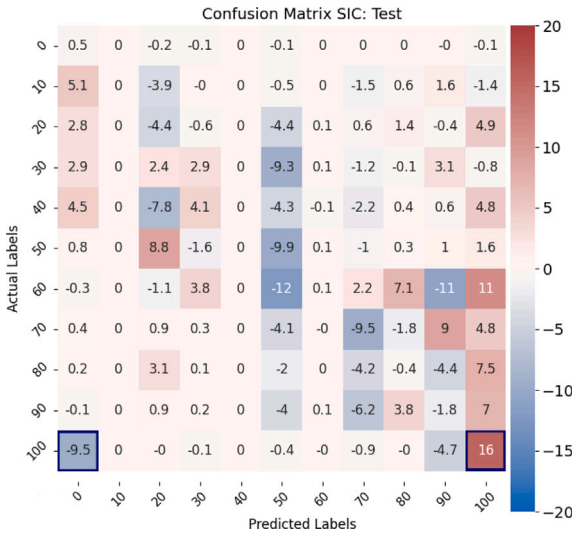


Fig. 17. The value change in each element of the SIC classification confusion matrix after adding ERA5 features, in comparison with the "HH, HV only" model.

References

Boulze, H., Korosov, A., Brajard, J., 2020. Classification of sea ice types in Sentinel-1 SAR data using convolutional neural networks. Remote Sens. 12 (13), 2165. http://dx.doi.org/10.3390/rs12132165.

Buus-Hinkler, J., Wulf, T., Stokholm, A.R., Korosov, A., Saldo, R., Pedersen, L.T., et al., 2022. AI4Arctic Sea Ice Challenge Dataset. http://dx.doi.org/10.11583/DTU.c.6244065.v2, Technical University of Denmark. Collection.

Chen, X., Patel, M., Cantu, F.J.P., Park, J., Turnes, J.N., Jiang, M., Xu, L., Scott, K.A., Clausi, D.A., Huang, W., 2023. The influence of input variable selection on deep learning-based sea ice parameter inversion from multi-sensor satellite data. In: OCEANS 2023-MTS/IEEE US Gulf Coast. IEEE, pp. 1–4. http://dx.doi.org/10.23919/OCEANS52994.2023.10337212.

Chen, X., Patel, M., Pena Cantu, F.J., Park, J., Noa Turnes, J., Xu, L., Scott, K.A., Clausi, D.A., 2024a. MMSeaIce: a collection of techniques for improving sea ice mapping with a multi-task model. Cryosphere 18 (4), 1621–1632. http://dx.doi.org/10.5194/tc-18-1621-2024, URL https://tc.copernicus.org/articles/18/1621/2024/.

Chen, X., Patel, M., Xu, L., Chen, Y., Scott, K.A., Clausi, D.A., 2024b. Weakly supervised learning for pixel-level sea ice concentration extraction using AI4Arctic sea ice challenge dataset. IEEE Geosci. Remote Sens. Lett. 21, 1–5. http://dx.doi.org/10.1109/LGRS.2023.3338061.

Cho, K., Naoki, K., 2020. Evaluation of AMSR2 thin ice area extraction algorithm applied to the sea ice zones of the Northern Hemisphere. In: 40th Asian Conference on Remote Sensing, ACRS 2019: Progress of Remote Sensing Technology for Smart Future. Acrs, pp. 1–10.

Comiso, J.C., Nishio, F., 2008. Trends in the sea ice cover using enhanced and compatible AMSR-E, SSM/I, and SMMR data. J. Geophys. Res.: Oceans 113 (C2), http://dx.doi.org/10.1029/2007JC004257, URL https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2007JC004257.

Day, J.J., Keeley, S., Arduini, G., Magnusson, L., Mogensen, K., Rodwell, M., Sandu, I., Tietsche, S., 2022. Benefits and challenges of dynamic sea ice for weather forecasts. Weather Clim. Dyn. 3 (3), 713–731. http://dx.doi.org/10.5194/wcd-3-713-2022, URL https://wcd.copernicus.org/articles/3/713/2022/.

De Gelis, I., Colin, A., Longépé, N., 2021. Prediction of categorized sea ice concentration from Sentinel-1 SAR images based on a fully convolutional network. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 5831–5841. http://dx.doi.org/10.1109/JSTARS.2021.3074068.

Gabarró, C., Hughes, N., Wilkinson, J., Bertino, L., Bracher, A., Diehl, T., Dierking, W., Gonzalez-Gambau, V., Lavergne, T., Madurell, T., et al., 2023. Improving satellite-based monitoring of the polar regions: Identification of research and capacity gaps. Front. Remote Sens. 4, 952091.

Gupta, B.C., Guttman, I., 2014. Statistics and Probability with Applications for Engineers and Scientists. John Wiley & Sons.

Han, H., Lee, S., Kim, H.-C., Kim, M., 2021a. Retrieval of summer sea ice concentration in the Pacific Arctic Ocean from AMSR2 observations and numerical weather data using random forest regression. Remote Sens. 13 (12).

Han, Y., Liu, Y., Hong, Z., Zhang, Y., Yang, S., Wang, J., 2021b. Sea ice image classification based on heterogeneous data fusion and deep learning. Remote Sens. 13 (4), 592. http://dx.doi.org/10.3390/rs13040592.

Holland, M.M., Bitz, C.M., 2003. Polar amplification of climate change in coupled models. Clim. Dyn. 21 (3–4), 221–232. http://dx.doi.org/10.1007/s00382-003-0332-6.

Imaoka, K., Kachi, M., Fujii, H., Murakami, H., Hori, M., Ono, A., Igarashi, T., Nakagawa, K., Oki, T., Honda, Y., Shimoda, H., 2010. Global change observation mission (GCOM) for monitoring carbon, water cycles, and climate change. Proc. IEEE 98 (5), 717–734. http://dx.doi.org/10.1109/JPROC.2009.2036869.

Karvonen, J., 2017. Baltic sea ice concentration estimation using Sentinel-1 SAR and AMSR2 microwave radiometer data. IEEE Trans. Geosci. Remote Sens. 55 (5), 2871–2883. http://dx.doi.org/10.1109/TGRS.2017.2655567.

Korosov, A., Demchev, D., Miranda, N., Franceschi, N., Park, J.-W., 2021. Thermal denoising of cross-polarized Sentinel-1 data in interferometric and extra wide swath modes. IEEE Trans. Geosci. Remote Sens. 60, 1–11.

Li, X.-M., Sun, Y., Zhang, Q., 2021. Extraction of sea ice cover by Sentinel-1 SAR based on support vector machine with unsupervised generation of training data. IEEE Trans. Geosci. Remote Sens. 59 (4), 3040–3053. http://dx.doi.org/10.1109/TGRS.2020.3007789.

Malmgren-Hansen, D., Pedersen, L.T., Nielsen, A.A., Kreiner, M.B., Saldo, R., Skriver, H., Lavelle, J., Buus-Hinkler, J., Krane, K.H., 2020. A convolutional neural network architecture for Sentinel-1 and AMSR2 data fusion. IEEE Trans. Geosci. Remote Sens. 59 (3), 1890–1902.

Malmgren-Hansen, D., Pedersen, L.T., Nielsen, A.A., Kreiner, M.B., Saldo, R., Skriver, H., Lavelle, J., Buus-Hinkler, J., Krane, K.H., 2021. A convolutional neural network architecture for Sentinel-1 and AMSR2 data fusion. IEEE Trans. Geosci. Remote Sens. 59 (3), 1890–1902. http://dx.doi.org/10.1109/TGRS.2020.3004539, URL https://ieeexplore.ieee.org/document/9133205/.

Melsom, A., Palerme, C., Müller, M., 2019. Validation metrics for ice edge position forecasts. Ocean Sci. 15 (3), 615–630. http://dx.doi.org/10.5194/os-15-615-2019, URL https://os.copernicus.org/articles/15/615/2019/.

Mudryk, L.R., Dawson, J., Howell, S.E.L., Derksen, C., Zagon, T.A., Brady, M., 2021. Impact of 1, 2 and 4 °C of global warming on ship navigation in the Canadian Arctic. Nature Clim. Change 11 (8), 673–679. http://dx.doi.org/10.1038/s41558-021-01087-6, https://www.nature.com/articles/s41558-021-01087-6.

Nagi, A.S., Kumar, D., Sola, D., Scott, K.A., 2021. RUF: Effective sea ice floe segmentation using end-to-end RES-UNET-CRF with dual loss. Remote Sens. 13 (13), 2460.

Nihashi, S., Ohshima, K.I., Tamura, T., 2024. Reconstruct the AMSR-e/2 thin ice thickness algorithm to create a long-term time series of sea-ice production in Antarctic coastal polynyas. Polar Sci. 39 (August 2023), 100978. http://dx.doi.org/10.1016/j.polar.2023.100978, https://linkinghub.elsevier.com/retrieve/pii/S1873965223000762.

Park, J.-W., Korosov, A.A., Babiker, M., Won, J.-S., Hansen, M.W., Kim, H.-C., 2020. Classification of sea ice types in Sentinel-1 synthetic aperture radar images. Cryosphere 14 (8), 2629–2645. http://dx.doi.org/10.5194/tc-14-2629-2020.

Radhakrishnan, K., Scott, K., Clausi, D., 2021. Sea ice concentration estimation: Using passive microwave and SAR data with a U-Net and curriculum learning. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 5339–5351.

Ren, Y., Li, X., Yang, X., Xu, H., 2021. Development of a dual-attention U-Net model for sea ice and open water classification on SAR images. IEEE Geosci. Remote Sens. Lett. http://dx.doi.org/10.1109/LGRS.2021.3058049.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. In: Intl. Conf. Med. Image Comput. Comput. Assist. Interv.. Springer, pp. 234–241. http://dx.doi.org/10.1007/978-3-319-24574-4_28.

Sandven, S., Spreen, G., Heygster, G., Girard-Ardhuin, F., Farrell, S.L., Dierking, W., Allard, R.A., 2023. Sea ice remote sensing—Recent developments in methods and climate data sets. Surv. Geophys. 1–37.

Scott, K.A., Buehner, M., Carrieres, T., 2014. An assessment of sea-ice thickness along the Labrador Coast from AMSR-E and MODIS data for operational data assimilation. IEEE Trans. Geosci. Remote Sens. 52 (5), 2726–2737. http://dx.doi.org/10.1109/TGRS.2013.2265091, URL http://ieeexplore.ieee.org/document/6553279/.

Stokholm, A.R., Buus-Hinkler, J., Wulf, T., Korosov, A., Saldo, R., Pedersen, L.T., Arthurs, D., Dragan, I., Modica, I., Pedro, J., Debien, A., Chen, X., Patel, M., Cantu, F.J.P., Turnes, J.N., Park, J., Xu, L., Scott, A.K., Clausi, D.A., Fang, Y., Jiang, M., Taleghanidoozdoozan, S., Brubacher, N.C., Soleymani, A., Gousseau, Z., Smaczny, M., Kowalski, P., Komorowski, J., Rijlaarsdam, D., van Rijn, J.N., Jakobsen, J., Rogers, M.S.J., Hughes, N., Zagon, T., Solberg, R., Longépé, N., Kreiner, M.B., 2023. The autoice challenge. EGUsphere 2023, 1–28. http://dx.doi.org/10.5194/egusphere-2023-2648, URL https://egusphere.copernicus.org/preprints/2023/egusphere-2023-2648/.

Stokholm, A., Wulf, T., Kucik, A., Saldo, R., Buus-Hinkler, J., Hvidegaard, S.M., 2022. AI4SeaIce: Toward solving ambiguous SAR textures in convolutional neural networks for automatic sea ice concentration charting. IEEE Trans. Geosci. Remote Sens. 60, 1–13.

Tamber, M.S., Scott, K.A., Pedersen, L.T., 2022. Accounting for label errors when training a convolutional neural network to estimate sea ice concentration using operational ice charts. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 15, 1502–1513.

Wang, L., Scott, K.A., Xu, L., Clausi, D.A., 2016. Sea ice concentration estimation during melt from dual-pol SAR scenes using deep convolutional neural networks: A case study. IEEE Trans. Geosci. Remote Sens. 54 (8), 4524–4533. http://dx.doi.org/10.1109/TGRS.2016.2543660.