Response to Reviewers for Paper

"MMSeaIce: Multi-task Mapping of Sea Ice Parameters from AI4Arctic Sea Ice Challenge Dataset"

(Chen, Patel, Pena Cantu, Park, Noa Turnes, Xu, Scott, Clausi)

***Note*:** the revised parts have also been highlighted with yellow color in the PDF version of the revised manuscript.

Response to Reviewers

Reviewer 1:

Reviewer Comment 1): *General Comments:*
This article reads extremely well, with no issues with the English language use and with good clear content.
I only suggest that the scientific novelties are presented with more emphasis than the competition for this scientific paper, rather than as an accidental consequence. The novel learnings and explanations are what makes this a scientific work, rather than just your method documentation. That said, there is far more novel content here than many recent ML submissions.
I conclude that this work is valuable and worthy, but should be revised to emphasise the scientific messages.

Response: The authors appreciate the reviewer's general comment. Specific comments have been addressed below.

Reviewer Comment 2): *Specific Comments:*
The mentioned imbalance of competition and scientific novelty is clear in the Abstract. The real science only appears in the very last sentence, with the showcasing of the various techniques, or components of the system, yet this is where the real scientific advancement lies. The abstract, and the rest of the paper, should summarise these messages and what we learned from the exercise.

Response: This is a very good suggestion. To emphasize the scientific advancement made by this research, the following sentences have been added after L8 in the abstract of the revised manuscript

> "Notably, the result analysis and ablation studies demonstrate that instead of model architecture design, a collection of strategies/techniques we employed lead to substantial enhancement in accuracy, efficiency, and robustness within the realm of deep learning-based sea ice mapping. Those techniques include input SAR variable downscaling, input feature selection, spatial-temporal encoding, and the choice of loss functions. By highlighting the various techniques employed and their impacts, we aim to underscore the scientific advancements achieved in our methodology."

Besides, the sentence in L195 has been revised as

> "Through comparison, the capability of those strategies in enhancing model performance is validated, as illustrated below."

In addition, the following sentence has been added in L261 of the revised manuscript to summarize the scientific advancement made in this research

> "In particular, we implemented several tricks to improve model accuracy, efficiency, and robustness. The techniques behind those tricks include input downscaling, feature selection, incorporating spatial and temporal information, and loss function design."

Reviewer Comment 3): *In terms of science, the method should also explain why you designed the network as you did, and why you designed the ablation study as you did? How does it characterise the significance of the different components? Were the components included or developed with certain expectations, e.g., have they been used before in different contexts perhaps? This is where we can learn the most about your method and the importance of various components.*

Response: The scientific reasons for designing the model architecture based on U-Net are as follows. First, the U-Net architecture is characterized by a U-shaped structure, with a contracting path (encoder) followed by an expansive path (decoder). This design allows the network to capture both high-level contextual information and fine-grained details. Besides, U-Net incorporates skip connections that connect the encoder and decoder at multiple resolutions. These connections enable the model to reuse feature maps from the encoding stage during decoding, helping to preserve spatial information and mitigate the vanishing gradient problem. In addition, U-Net has demonstrated effectiveness, especially in scenarios with limited annotated data. The architecture's ability to learn from small datasets and generalize well to new data is crucial in practical applications where acquiring extensive labeled data is challenging, such as sea ice mapping. Thus, the following sentences have been added after L95 in the revised manuscript

> "The network designed in this research is based on the architecture of a U-Net due to the following reasons. Characterized by the U-shaped structure, the network is able to capture both high-level contextual information and fine-grained details. Besides, the incorporation of skip connections facilitates the reuse of feature maps, addressing spatial information preservation and the vanishing gradient problem. Moreover, U-Net's demonstrated efficacy, particularly in scenarios with limited annotated data like sea ice mapping, underscores its ability to learn effectively from small datasets and generalize to new, challenging data environments."

Ablation studies are essential in scientific investigations involving deep learning models to systematically assess and understand the impact of various components or factors, such as different data inputs in this research. These studies help identify the specific contributions and importance of each input, allowing researchers to pinpoint which elements significantly influence model performance. In the context of our study,

conducting ablation studies on different data inputs enables a nuanced examination of their individual effects on the model's ability to accurately predict sea ice characteristics. This scientific approach aids in unraveling the intricate relationships between input features and model outcomes, guiding the optimization of model architectures and data preprocessing techniques for improved performance and interpretability. Thus, the following sentences have been added after L188 in the revised manuscript

> "In the context of our study, conducting ablation studies on different data inputs enables a nuanced examination of their individual effects on the model's ability to accurately predict sea ice characteristics. This scientific approach aids in unraveling the intricate relationships between input features and model outcomes, guiding the optimization of model architectures and data preprocessing techniques for improved performance and interpretability."

Yes, the components included or developed in this research have been used before in different contexts. For example, the SAR image downscaling operation was implemented in a previous work (Liu et al., 2021, doi: 10.1109/JSTARS.2021.3122546) concerning sea ice classification to avoid the appearance of scalloping and interscan banding artifacts in classification results. The incorporation of spatial and temporal information as data inputs originates from a previous work concerning the sea ice thickness estimation with Google Earth Engine and Sentinel-1 GRD data (Shamshiri et al., 2021, doi: 10.1016/j.rse.2021.112851). As for loss function selection, in a recent study by Kucik et al. (Kucik and Stokholm 2023, doi: 10.1038/s41598-023-32467-x), a U-Net architecture was trained on the AI4Arctic dataset to accurately retrieve sea ice concentration (SIC) with different loss functions for performance comparison.

Therefore, the following sentence has been added in L123 in the revised manuscript

> "This downscaling operation has also been implemented in a previous work (Liu et al. 2021a) concerning sea ice classification to avoid the appearance of scalloping and interscan banding artifacts in classification results."

Besides, the following sentence has been added in L150 in the revised manuscript

> "The incorporation of spatial and temporal information as data inputs originates from a previous work concerning the sea ice thickness estimation with Google Earth Engine and Sentinel-1 GRD data (Shamshiri et al. 2021)"

In addition, the sentence L102 has been revised as

> "For example, in a recent study by Kucik et al. (Kucik and Stokholm 2023), a U-Net architecture was trained on the AI4Arctic Sea Ice Dataset version 2 (ASID-v2) (Saldo et al. 2021) to accurately retrieve SIC with different loss functions for performance comparison."

Reviewer Comment 4): *Consider whether the title can somehow reflect that the science is somehow this contribution/significance analysis of the components. Might be difficult and is not critical though.*

Response: As suggested, the title of this paper has been revised as "MMSeaIce: a Collection of Techniques for Improving Sea Ice Mapping with a Multi-task Model" in the revised manuscript.

Reviewer Comment 5): *I suggest that you add sub-headings on the left with the different models to explain what they are and their relation to the ablation study and table 4. That is, remind the viewer which is the "full model", that model 2 has "no downscaling", and model 8 uses "cross-entropy", etc. This would make it easier to try to interpret the causes of the resultsF.*

Response: We agree. The suggested subheadings have been added to all the bulletin points (Page 10) in the revised manuscript.

Response to Reviewers

Reviewer 2:

Reviewer Comment 1): *First of all, I would like to congratulate the authors on their first place in the competition and a well-written, concise and informative report of their findings.*

Broad Comments:

From a technical standpoint I find the manuscript to be well constructed and easy to follow. I do believe some extra discussion would benefit the work and help place it into the greater context of the ongoing efforts of sea ice classification in a changing Arctic.

Response: The authors appreciate the reviewer's broad comment. Specific comments have been addressed below.

Reviewer Comment 2): *The two things I would like to see discussed in additional detail would be:*

The influence of the ice charts as ground truth in terms of what the resulting classifier is capable of extracting and what is outside of the scope of classification. In the introduction some of the uses of ice charts and the multitude of output variables is mentioned; it seems to me that a regional sea ice concentration and floe size as is predicted here, could be derived from a classified map if the classification took place at the same effective resolution as the SAR sensor, for example.

Response: The ice charts contain the ice characteristics of each polygon, including a total ice concentration for the polygon (CT), concentrations of up to 3 different ice stages of development/floe sizes (forms) specified by their partial concentration (CA, CB and CC), their stage of development (SA, SB and SC) and their floe size/form (FA, FB and FC). Thus, the utilization of ice charts as ground truths enables the classifier to extract sea ice concentration (SIC), stage of development (SOD), and floe size (FLOE) at region level. Although pixel-based labels produced from the ice charts are provided in the ready-to-train version of the AI4Arctic dataset, they are generated based on a thresholding approach and cannot tell us about the locations of different ice types/floe sizes at SAR sensor resolution. That being said, the extraction of the sea ice parameters mentioned above at SAR sensor resolution is out of the scope of classification in this research. Besides, some other ice characteristics, such thickness and drift, are also outside the scope of this research due to a lack of such information in the ice charts. Therefore, to discuss this, the following sentences have been added after L87 in the revised manuscript

> "The utilization of ice charts as ground truths enables the classifier to extract the three sea ice parameters mentioned above at region level. Although pixel-based labels produced from the ice charts are provided in the ready-to-train version of the AI4Arctic dataset, they are generated based on a thresholding approach and cannot tell us about the locations of different ice types/floe sizes at SAR sensor resolution. That being said, the extraction of the

sea ice parameters mentioned above at SAR sensor resolution is out of the scope of classification in this research. Besides, some other ice characteristics, such thickness and drift, are also outside the scope of this research due to a lack of such information in the ice charts."

Reviewer Comment 3): *The effect of including time and spatial information in the classification and what that might mean for using such a classifier in a changing Arctic. In a wider scope, one could ask the question if there might be a conflict between performing best on historical data and performing best in an uncertain future. This can be discussed in terms of which input variables are used, how the class imbalance is handled, etc.*

Response: This is a very good suggestion. The inclusion of temporal and spatial information signifies the integration of sea ice climatology knowledge into the classification process. While this enhancement demonstrates improved model performance on recent data, it is essential to acknowledge the inherent limitations of relying solely on climatological information. The dynamic nature of the Arctic, undergoing continuous changes, emphasizes the continued reliance on observations from diverse sensors, such as SAR and passive microwave, ensuring that satellite data occupies a predominant role in the input channels for robust sea ice mapping. To discuss this, the following sentences have been added after L215 in the revised manuscript

"The inclusion of temporal and spatial information signifies the integration of sea ice climatology knowledge into the classification process. While this enhancement demonstrates improved model performance on recent data, it is essential to acknowledge the inherent limitations of relying solely on climatological information. The dynamic nature of the Arctic, undergoing continuous changes, emphasizes the continued reliance on observations from diverse sensors, such as SAR and passive microwave, ensuring that satellite data occupies a predominant role in the input channels for robust sea ice mapping."

Reviewer Comment 4): *Specific Comments:*

L.10: The authors claim that the tested techniques significantly improve the robustness of models, is this a qualitative finding or is there some quantitative analysis backing up this statement? Maybe this is unclear because robustness is not uniquely defined in this context.

Response: We acknowledge that the "robustness" here is not uniquely defined. What we would like to mention here is that our model has shown relatively high stability with low variations, as demonstrated in the "standard deviation" columns in Table 4 of the manuscript. To clarify this, the word "robustness" has been removed.

Reviewer Comment 5): *Sec 3: I am sorry if I just missed it, but I would like some discussion on input data preparation. I assume that some of the auxiliary data was brought up to input patch dimensions and added as*

channels because of convenience, but might this have an effect on the classifier (e.g. vs adding them in the bottleneck)?

Response: Yes, the auxiliary data are brought up to input patch dimensions and added as channels in this research. Although it is also feasible to add them in the bottleneck, adding them as input channels facilitates us to analyze the effect of choosing different data inputs on model performance. Besides, it enables the CNN model to extract pixel-based nonlinear features at the very beginning. Nevertheless, in future works it would be interesting to compare the current channel adding approach vs adding them in the bottleneck. To discuss this, the following sentences have been added after L137 in the revised manuscript

> "The auxiliary data are brought up to input patch dimensions and added as channels in this research. Although it is also feasible to add them in the bottleneck, adding them as input channels facilitates us to analyze the effect of choosing different data inputs on model performance. Besides, it enables the CNN model to extract pixel-based nonlinear features at the very beginning. Nevertheless, in future works it would be interesting to compare the current channel adding approach vs adding them in the bottleneck."

Reviewer Comment 6): *L.129: Why were the months discretized for input instead of a continuous approach and what are the possible implications for the classification?*

Response: The reason to discretize time information instead of using continuous values (i.e., values specific to day) is that since the ice climatology is similar within one month, adopting continuous values might not improve model performance significantly. Besides, the imbalanced data distribution between different dates might lead to overfitting. In contrast, the data volume available for each month is relatively balanced. In future works, when the next version of the dataset is released (with around 16 times more data), it would be interesting to adopt the continuous approach for comparison. To clarify this, the following sentences have been added after L152 in the revised version

> "The reason to discretize time information instead of using continuous values (i.e., values specific to day) is that since the ice climatology is similar within one month, adopting continuous values might not improve model performance significantly. Besides, the imbalanced data distribution between different dates might lead to overfitting. In contrast, the data volume available for each month is relatively balanced. In future works, when the next version of the dataset is released (with around 16 times more data), it would be interesting to adopt the continuous approach for comparison."

Reviewer Comment 7): L.197: The predictions aren't really 'polygon based' are they? Maybe spatially smoothed predictions or some similar wording might be more fitting.

Response: Sorry for the confusion. In the revised manuscript, the "polygon-based" has been replaced by "spatially smoothed".

Reviewer Comment 8): L.197-200: Some published methods exist that make use of various input scales, maybe this could be mentioned/referenced here.

Response: As suggested, the following sentence has been added after L123 to mention the works that make use of various input scales for sea ice mapping.

> "Various input scales have also been implemented in a previous work (Stockholm et al. 2022) concerning sea ice concentration estimation."

Also, The paper cited above have been added as new Reference

Stokholm, A., Wulf, T., Kucik, A., Saldo, R., Buus-Hinkler, J., and Hvidegaard, S. M.: AI4SeaIce: Toward Solving Ambiguous SAR Textures in Convolutional Neural Networks for Automatic Sea Ice Concentration Charting, IEEE Trans. Geosci. Remote Sens., 60, 1–13, 2022.

Response to Reviewers

Reviewer 3:

Reviewer Comment 1): *Review of the submitted manuscript, "MMSeaIce: Multi-task Mapping of Sea Ice Parameters from AI4Arctic Sea Ice Challenge Dataset". The manuscript investigates methods enabling the mapping of different sea ice parameters, which were applied in The AutoICE Challenge, achieving the first spot on the challenge podium. Thank you for a well-written manuscript covering interesting results from The AutoICE Challenge, including both additional information on the developed method but also additional tests post-competition that verify previous assumptions on key matters that are important for developing deep learning models to automatically map sea ice in the polar regions from, among others, SAR imagery. To summarise the comments, there are very few things that need clarification. Some comments have minor suggestions for grammatical corrections or rephrasing. A PDF with small grammatical suggestions is attached. The manuscript is of high quality, covering an important topic, and is recommended for publication after a minor revision.*

Response: The authors appreciate the reviewer's general comment. The suggested corrections in the attached PDF have been made in the revised manuscript. Other specific comments have been addressed below.

Reviewer Comment 2): *Title: I believe a the would be appropriate in the title so that it reads: "MMSeaIce: Multi-task Mapping of Sea Ice Parameters from the AI4Arctic Sea Ice Challenge Dataset"*

Response: We agree. As Reviewer 1 suggested, we have changed the title into "MMSeaIce: a Collection of Techniques for Improving Sea Ice Mapping with a Multi-task Model".

Reviewer Comment 3): *I suggest including a reference to the paper "The AutoICE Challenge" once it is available as a preprint in The Cryosphere, which should be very soon (it has been accepted but awaiting posting). A reference to this manuscript will be included in the "The AutoICE Challenge" article during the initial review phase.*

Response: Yes, the reference to the AutoICE Challenge paper has been added in the revised manuscript.

Reviewer Comment 4): *L5: I think "using Sentinel-1 SAR data" can be removed, as this is mentioned in the proceeding sentence.*

Response: The suggested correction has been made in the revised manuscript.

Reviewer Comment 5): *L32: This sentence is slightly negative. Instead of "it is important to acknowledge the limitations of previously proposed DL-based methods", I suggest writing: "it is important to acknowledge potential areas for improvements of previous proposed DL-based methods" or similar.*

In connection to this on L34, there are also advantages in utilising a singular sensor type, which, among others, simplifies operational aspects and can enable the investigation of how to extract its maximum value.

Response: As suggested, the sentence in L34 has been revised as

> "However, it is important to acknowledge potential areas for improvements of previous proposed DL-based methods."

Also, the sentence in L36 has been revised as

> "Although this simplifies operational aspects and can enable the investigation of how to extract its maximum value, it might lead to potential ambiguities and limitations in information integration."

Reviewer Comment 6): *L48: I suggest you write the abbreviations for SIC, SOD and FLOE here instead of L68.*

Response: The suggested correction has been made in the revised manuscript.

Reviewer Comment 7): *L54: "a bag of tricks" -> "a collection of strategies/techniques" /*

Response: The suggested correction has been made in the revised manuscript.

Reviewer Comment 8): *L92: The AI4Arctic dataset in question here is the "AI4Arctic Sea Ice Dataset version 2"(ASID-v2). I think you should add a reference here to. https://data.dtu.dk/articles/dataset/AI4Arctic_ASIP_Sea_Ice_Dataset_-_version_2/13011134*

Response: The suggested reference has been added in the revised manuscript.

Reviewer Comment 9): *L104: I think it is important to acknowledge that the ice analysts do not willingly create ice charts in low resolution but rather that it is for the sake of efficiency and lack of time instead of inability. Suggestion: "who have to produce charts in low resolution due to time constraints", or something similar.*

Response: We agree. The sentence has been revised in L116 of the revised manuscript as

> "Despite the resolution of the SAR imagery that is well suited for SAR sea ice monitoring, the polygon egg code data is derived from the knowledge of ice analysts who have to produce charts in low resolution due to time constraints."

Reviewer Comment 10): *Table 2: "produces the highest accuracy", perhaps it is the highest combined score, which is referred to?*

Response: Yes. To clarify this, the word “accuracy” has been replaced with “combined score”.

Reviewer Comment 11): *Table 4: The accuracy scores you report, is it actually the accuracy or instead the default scores for SIC, SOD and FLOE, i.e. R2, F1, F1, respectively? This could be clearer.*

Response: Yes. To clarify this, the first sentence of the caption of Table 4 has been revised as

> “The average default scores for SIC, SOD and FLOE (i.e. R2, F1, F1) obtained from models with different configurations.”

Reviewer Comment 12): *L199: A larger patch size does not lead to a larger receptive field in itself. Theoretically, at least, instead, it should allow for training models, which have a larger receptive field effectively. An explanation for not reaching higher scores with larger patch sizes could instead be a consequence of utilising a model with an insufficient receptive field for the patch size. I think this sentence should be revised. Furthermore, I think this could be an area for further improvements to the model, which you could consider adding a sentence about.*

Response: We agree. To clarify this, the following sentence has been added after L240 in the revised manuscript

> “This could be due to a consequence of utilising a model with an insufficient receptive field for the patch size, which could be an area for further improvements to the model in future works.”

Reviewer Comment 13): *Tables in general: I think you should consider adding some more text to describe what is in the tables, as this is very minimal in the current manuscript.*

Response: As suggested, we have added some more text in the caption of each table to have a better description of the content inside.