Review of "Data-driven emulation of melt ponds on Arctic sea ice" by Simon Driscoll1 et al.

Summary:

The positive feedback effect of melt ponds accelerates the Arctic sea ice melting process, significantly impacting the sea ice mass balance. However, statistically based physical parameterization schemes still exhibit significant uncertainty in representing subgrid-scale melt pond evolutions. In this study, the authors trained a machine learning emulator based on satellite observations as an alternative to the melt pond parameterization in a column sea ice thermodynamic model (i.e., Icepack). They emphasized that this emulator has the potential to be further integrated into climate models.

Overall, this is timely work, as the melt pond fraction on Arctic sea ice is increasing, and accurately simulating melt ponds can reduce the uncertainty of future climate projections in climate models. Regrettably, however, I find the current manuscript is not as well prepared as it should be for submission; the text makes for uncomfortable reading. Especially since the method is poorly introduced, does not convince the importance of its use to conclude, and does not meet the quality and reputation of The Cryosphere, I have to recommend **rejecting** its publication. Please find my specific comments below.

Specific comments:

1. Methods

The description of the method is so confusing that I cannot discern how the authors trained the emulator. At least the following points hinder my understanding:

(1) The dataset splitting

The authors said that the training dataset is that the 2002-2011 MERIS data and the 2017-2019 OLCI data (lines 130 and 132), but the data in 2019 were also used in the model validation (line 174)?

(2) The feature data

As listed in Table 1, the DMIOI-L4 sea ice fraction and analysed ST (end on 31 May 2021) are the "features". If I understand correctly, it means they are the "inputs" for the emulator. Therefore, during training and testing, the emulator's inputs should be consistent. However, data for these two variables is only available up to May 31, 2021. Does this mean that after this date, up to 2022, the emulator did not use these two variables as inputs? Or were data from another dataset used instead?

(3) Training of the emulator

I commend the authors for using the Hyperband algorithm to automatically optimize the model's hyperparameters and obtain the best combination. However, I am confused as to why manual hyperparameter tuning was also performed. According to the description below (line 179), the authors ultimately employed a fully connected neural network with 10 hidden layers and 10 nodes per layer. This configuration seems to have been determined by the Hyperband algorithm rather than manual tuning. If the intention is to discuss the impact of different hyperparameters on the results, I

suggest moving this content to the discussion section.

(4) Applicability of the emulator

What surprises me is that the authors, for each grid point, fit the daily MPF using features such as daily 10-meter wind speed and other variables. This emulator, which can be understood as a simple "multivariate nonlinear regression" model, even if trained to be very realistic, cannot be proven as a substitute for parameterization schemes. The input-output scenarios of this emulator do not align with those in model parameterization schemes, which typically operate with shorter time steps. Therefore, I am not convinced the authors have provided sufficient evidence that this emulator is capable of replacing the model's parameterization schemes.

(5) Significance of mutual analysis

I did not fully understand why the authors calculated the mutual information scores for each feature in this section. All features were ultimately used to establish the emulator, even though some of them might be less important, right? This part of the analysis seems more appropriate for moving into the discussion section to enhance the interpretability of the emulator.

2. Data

I have strong concerns about the "version 1.5" MERIS and OLCI data shown here. I did not check the v1.5 dataset, but its updated version (Istomina et al., 2023) revealed an overall positive trend (+0.15% to +3% per decade) of the Arctic MPF (also can be seen in their Figures 9 and 12). I'm unsure if I missed something, but there indeed are some other observations that support the increasing Arctic MPF (e.g., Feng et al., 2022; Xiong and Ren, 2023), contrasting with Figure 1 in this manuscript. I strongly suggest that the authors check for errors in the way they handle the data.

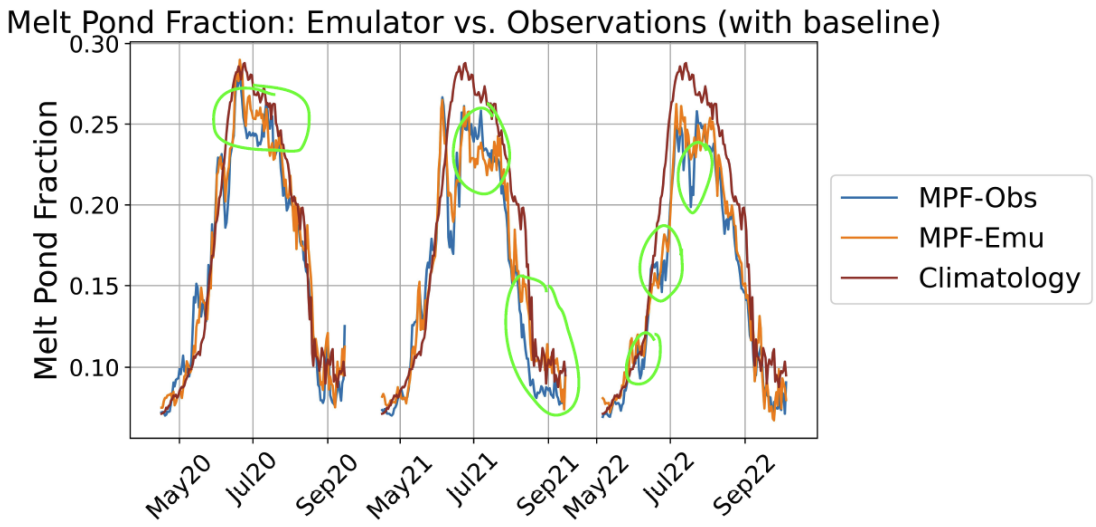
3. Results

Section 3 presents results in a structurally disorganized manner. A more coherent approach would be to present the model validation results first, followed by the test results.

The current presented results have at least two critical shortcomings: the performance and validation of the emulator. The authors described "the emulator shows a very strong similarity to the observed MPF" in line 184, however, from my perspective, although the emulator shows overall similarity with observations in the test set, there are many obvious mismatches (see areas circled in green in the figure below). Thus, the authors' description of the results is imprecise and overly colloquial. On the other hand, I believe the current validation approach is ineffective. It is necessary to compare the emulator with the original parameterization scheme on the same test set, rather than only comparing it against the climatology (lines 190-192). In other words, I believe the authors have not achieved their stated goal of replacing the physical parameterization scheme with the emulator (also refer to my fourth comment on the methods).

Furthermore, I suggest that the authors validate the emulator's effectiveness using MOSAiC data (e.g., Webster et al., 2022) and compare it with the original

parameterization scheme, which would render the study more comprehensive.



Other issues:

Please note that because there are numerous language and formatting issues in this manuscript, only several of them are listed below. To improve the quality of your manuscript, I recommend thoroughly revising the language to ensure a smoother flow and clarity.

- The language of this paper is excessively verbose and lacks academic rigor. I mean, one should not use vague terms such as "very" to describe results (e.g., line 60, line 62, line 66, line 184, line 244).
- Figure 1: Which line represents the "emulator" mentioned in the title?
- Figure 3: I do not think this simple training workflow worth a schematic figure to illustrate. The only information I can get from this schematic figure is that the authors interpolated the features onto a widely used polar stereographic projection grid.
- Lines 187-189: I am unsure if the emulator has not seen this "large scale refreezing" in the test dataset, and thus, I am not convinced by this statement.
- Lines 228-229: What does this sentence mean? Not clear.
- The formatting of the references is highly messy.

References:

Feng, J., Zhang, Y., Cheng, Q., and Tsou, J. Y.: Pan-Arctic melt pond fraction trend, variability, and contribution to sea ice changes, Global and Planetary Change, 217, 103932, https://doi.org/10.1016/j.gloplacha.2022.103932, 2022.

Istomina, L., Niehaus, H., and Spreen, G.: Updated Arctic melt pond fraction dataset and trends 2002–2023 using ENVISAT and Sentinel-3 remote sensing data, https://doi.org/10.5194/tc-2023-142, 22 September 2023.

Webster, M. A., Holland, M., Wright, N. C., Hendricks, S., Hutter, N., Itkin, P., Light, B., Linhardt, F., Perovich, D. K., Raphael, I. A., Smith, M. M., Von Albedyll, L., and Zhang, J.: Spatiotemporal evolution of melt ponds on Arctic sea ice,

Elementa: Science of the Anthropocene, 10, 000072, https://doi.org/10.1525/elementa.2021.000072, 2022.

Xiong, C. and Ren, Y.: Arctic sea ice melt pond fraction in 2000–2021 derived by dynamic pixel spectral unmixing of MODIS images, ISPRS Journal of Photogrammetry and Remote Sensing, 197, 181–198, https://doi.org/10.1016/j.isprsjprs.2023.01.023, 2023.