Containing Author Comments

Review 1

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.

We thank the reviewer for their insights. A large effort with our co-authors has been made to substantially improve the clarity of our text and remove confusion we caused. Aided by the reviewer's insights, we have implemented notable scientific changes. We believe all together these have made a substantially improved paper that addresses the points made. We hope to convince the reviewer with the following answers and are incredibly grateful for the reviewer's input that led to these improvements.

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)?

We thank the reviewer for highlighting this and apologise for the confusion we caused. In machine learning, when the datasets are short it is common practice to use all validation and training data for final training, whereas test data is kept separately. This is particularly useful for the cryosphere, where data is often sparse or missing. We also refer the reviewer to point 3, where we have much more clearly clarified the minimal role hyperparameter tuning played and have moved it to the conclusions section in our paper as per the suggestion by the reviewer. In this section we then fully clarify our approach.

(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?

This is a copy paste error. When originally looking into datasets many years ago the DMI had datasets that ran up to this time, now seamlessly extended. We use the full same dataset that covers the whole period, and this has been updated in our table. We thank the reviewer for spotting this error for us.

(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.

Thank you for this suggestion which has made us improve the description of our approach. We have not ultimately used automatic hyperparameter tuning in our choice of final model. We did this manually and have clarified our wording which was confusing - hyperparameter optimisation of complex models was ultimately not necessary as simpler smaller models were equally good. As well as clarifying what we did we have moved this portion of the text 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.

We thank the reviewer for this comment. We have clarified that our objective is not to show that such an emulator can replace the model's parametrisation schemes. Nor do we aim to emulate the full physical complexity of the evolution and formation of melt ponds. Instead, we seek to identify useful observational variables and learn the relationship between observed inputs and melt pond fraction – which is understood here as a proxy for albedo on sea ice and thus is a key variable relevant to the Arctic's energy budget and is integral in models. By showing that a relationship exists between key observed variables, and particularly those often used in models, we believe this is an important step – as well as informative for the physical understanding of melt ponds.

We have discussed that showing this relationship exists is valuable in itself - going from emulators based on synthetic/model data to those based on noisy, sometimes missing, observational data has been a difficult, major and often prohibitive step in many areas of climate research. We expand on the significance of the research. The design choices that we made for simplification (no advection, daily averaged values) might be prohibitive for replacing a model parametrisation, and we now discuss this in more detail, but our study makes us optimistic about the future steps for data-driven developments in the field. Our manuscript title has been updated to highlight we are solely targeting melt pond fraction. We have added text to highlight the challenges that such emulators might have in replacing model parts, such as advection and time stepping issues, as well as the complex processes involved in melt pond formation, and thus in physical models. In doing so we also consider where data-driven methods might be useful in aiding future scientific inquiry and enriching model processes - such as in identifying shifts in fractal dimensions and interconnectivity in melt ponds from imagery.

(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.

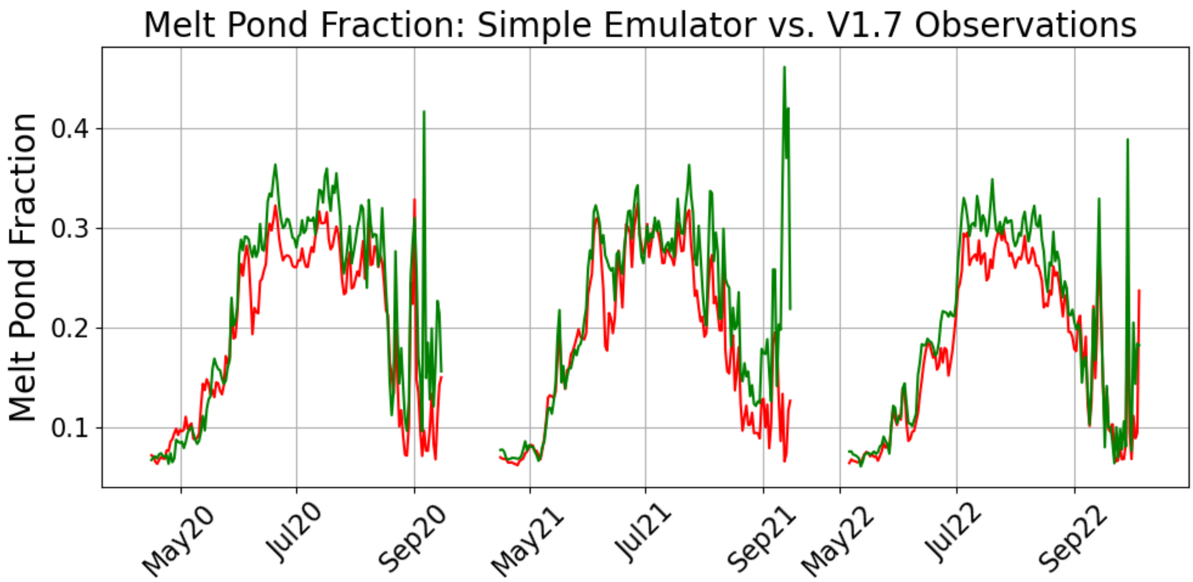
We thank the reviewer for their useful comment, and your interpretation is correct. We have indeed moved the mutual information section to the conclusions to enhance the interpretability of the emulator.

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.

We thank the reviewer for their valuable insights on this as it has improved our paper. We can confirm the data handling to produce Figure 1 is indeed very straightforward (using built in python functions), and that the step change drop between the MERIS and OLCI is a real feature, but a limitation, of using the "V1.5" dataset. We thank the reviewer for pointing this out as we now have also tested our approach on the updated V1.7 dataset (Istomina et al. 2023) where we also observe the values as seen in this latest Istomina paper and we do not observe the step/decline present in V1.5.

We are very grateful to the reviewer for this suggestion, as by applying our pipeline to this V1.7 dataset also, we are not only able to successfully emulate melt pond fraction on the newer dataset, but our skill scores also improve. We explain this as being due to a lack of step change disparity between satellite products in this updated dataset. Furthermore, models here can be simpler which aids interpretability of melt pond processes and our physical understanding. We provide an example below for the V1.7 dataset: a simple emulator (red) versus observations (green), trained on V1.7 training data (not shown) and evaluated on the V1.7 test data portion (shown). We now include such examples in our manuscript.



We have therefore now noted in our paper the limitations of using the V1.5 dataset alongside the success of this approach by testing it on the latest V1.7 dataset. We furthermore note this highlights the fact that our approach and pipeline is robust and general across different datasets – valuable for future data releases.

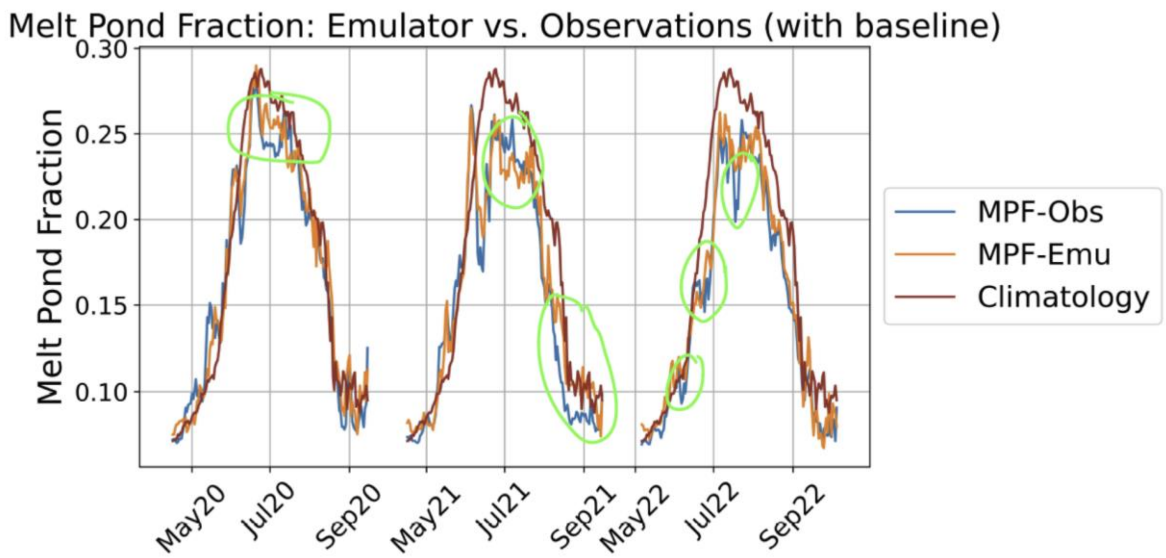
We are very grateful to the reviewer and the extra work it led to. By showing success on the newer Istomina et al. (2023) V1.7 dataset, that the reviewer guided us to, we feel it has improved our manuscript substantially.

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.

We thank the reviewer for this useful comment. We have organised our results in a more structured manner which aids clarity and a better narrative structure.

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.



We thank the reviewer for pointing this out. We have added extra discussion around the points the reviewer has highlighted. We also note that no emulator made from observations (which are noisy, sometimes missing and so on) will ever be a perfect replica of the physical phenomena and thus note the limitations they may have.

As to the second point, we apologise for the confusion caused which is a key misunderstanding due to our text. We do not involve a physical parametrisation in our paper, and it is not our aim to replace a physical parametrisation in our paper. We note that these developments could indeed aid data-driven components in the future. The direct goal of our paper is to show that a pointwise emulator can predict observed melt pond fraction from other observational variables and, therefore we do not have a physical parametrisation to work with here. We have clarified that this is a notable step for many areas in climate research which struggle to go from model data to observational data, and we have expanded on this value, but we are clearer now about the many challenges one might face before such methods are able to replace model parametrisation schemes, which is not our goal. We have hopefully clarified this in the text.

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.

We thank the reviewer for these insights and address both the specific comments below and have gone through this more thoroughly in general.

- 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).

Thank you for this, we have removed these terms and tightened our language. We have also gone through the text to make it sound more professional.

- Figure 1: Which line represents the "emulator" mentioned in the title?

This a typo. We apologise and have amended 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.

When presenting to audiences that were not familiar with machine learning they found this schematic helpful to visualise an ML training pipeline and process. In order to be inclusive to audiences we have kept the schematic for now, but we are happy to be flexible.

- 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.

We agree that suggesting it is a large-scale refreezing is ambitious and have gone with the much more modest suggestion that it might capture physical events in the Arctic associated with weather variability.

- Lines 228-229: What does this sentence mean? Not clear.

We mean would anticipate an emulator trained on observations ideally then matches better the observed melt onset – because it is on that dataset on which it is essentially trained. Thus, we have added “For example, an emulator trained on observations inherently captures the timing of melt onset as recorded in those observations". We have clarified this adding more detail.

- The formatting of the references is highly messy.

We thank the reviewer for pointing this out, some of which came from the bibtex. We have reformatted the references extensively to be neat and consistent.

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.