Authors' response to interactive comments by Anonymous Referee #2

January 26, 2021

Black text: Reviewer's comment

Blue text: Authors' response

1 General Comments

This manuscript describes the application of machine learning techniques, specifically an ensemble of multilayer perceptrons, to estimate the hydrological variable Snow Water Equivalent (SWE). As described by the authors, SWE is a crucial variable which is difficult to directly measure at scale. The paper subject advances the application of ML techniques for SWE estimation by application of ensemble methods, discerning model applicability over climatic regions. and demonstrating advantages over empirical methods.
This research builds on the Ordy et al (2020) study by extending the geographical scope, using direct estimation of SWE and introducing snow classes in MLP training. There is sufficient novelty in this paper for publication, however it should first be strengthened in clear justification for decisions, interpretation of results and conclusion.
The manuscript leaves out some essential elements from Ordy et al, including descriptions of evaluation metrics and why they are selected. The addition of explanatory variables such as Snow Density from ERA5 lack explanation and background. Following strong results reporting sections, the discussion finding and conclusions of the manuscript require additional reflection on the limitations of the study and the context.
Overall, a strong effort worthy of publication on the basis of some revision and structural improvement. Some coherence is missing in the experimental design, in the inclusion of variables, the applicability of the study and the conclusions drawn from it. These require revision, hope the comments that follow can be of help.

We thank the reviewer for their comments and we appreciate the effort put in the revision. We will take their comments into account to strengthen the output of the study. In the following, we address each specific comment and explain how we will incorporate them into the manuscript.

2 Specific Comments

1. Pg 1, ln 15: ”Using a greater number of MLP parameters could lead to further improvements” It is somewhat self-evident that increased parameterization of an MLP model could potentially produce better results, can this statement be focused to the study specific outcomes?

   We will summarize the outcomes of the study more precisely. Specifically, we will focus on using SWE instead of density as the target variable, testing several options of ANN structural characteristics (e.g. optimization algorithm, activation function, parameter initialization, increasing the number of parameters) improves estimates of SWE, including input uncertainty on snow depth improves the model's performance and dividing the area into snow classes and using an individual ANN model for each of them gives a greater representation of the geophysical diversity of snow.

2. pg 2, 50: This description of the application of physics-based models for SWE estimation is a bit too simplistic here, given ERA5 snow density as used later as an explanatory variable. The iSnobal mentioned is a coupled energy and mass-balance model that requires a great deal of meteorological data derive accumulating snow density and in turn modelled SWE. Please provide some further description on the advanced requirements these approaches and limitations, beyond only computational cost

   Physical-based models like iSnobal take input variables (e.g. incoming longwave radiation, soil temperature, net solar radiation) which are not available in real time in Canada. Furthermore, physical-based models are, as Painter et al. [2016] mentioned, the logical choice for distributed SWE estimates. However, we aim for a conversion model based on data points which are sparely scattered in time and space and uses only variables available in real time. We will add these thoughts in the revised manuscript. Further, ERA5 snow density was included as a test of reanalysis data, but is not available in real-time. Further discussion is given in comment 13.

3. ln 54: Consider including Snauffer et al, 2018. https://doi.org/10.5194/tc-12-891-2018.

   We would like to thank the reviewer for mentioning related literature. We will include this article into the mentioned section.

4. pg 3, ln 76: The second hypotheses seems too broad. ”in-depth testing”. Please be more specific as to the methodology to be tested.

We do not want to change the hypothesis, because they have been determined before the study. However, we will be more specific with the outcomes of the study in this paragraph, in accordance with the answer to comment 1.

5. pg 3, ln 78: ”The entire area of Canada.” Is this an overreach given the the limited density of measurements across much of Canada? ”Applicability in a broader context”. Be more specific in the what this broader context is.

   It was not our intention to overreach, but we wanted to mention that the data set is scattered sparsely and non-uniformly over the entire area of Canada, which is very large. Testing the ”applicability in a broader context” means that the data set includes almost all the Sturm’s snow classes (except the ice snow class, for which there is too little data to analyze), which gives the opportunity to test the model’s applicability to multiple snow class zones. We will refine the explanations in the revised version of the manuscript accordingly.

6. pg 3, ln 78: This last sentence seems out of place to close the paragraph. Moving one sentence earlier would improve the paragraph.

   In accordance with your comment, this sentence will be revised and probably linked together into one sentence.

7. pg 3, ln 94: Is MSE the definitive objective function for regression problems? Better likely to phrase as ”commonly used”

   Following the mathematical theory by Goodfellow et al. [2016], the MSE is derived from the the maximum likelihood estimator when dealing with regression models and therefore, the best choice. However, ”commonly” can be included, since the determination of the objective function is ultimately the modeller’s choice.

8. pg 5, ln 131: ”The algorithms RMSProp and AdaDelta produce good results”. Please elaborate in this statement, or tie in better with the following two sentences.

   We are sorry that the formulation is misleading. The statement is indeed linked to the previous sentence. Schaul et al. [2014] compared stochastic gradient methods by testing them on small-scale problems, and concluded that the algorithms RMSProp and AdaDelta produced good results. This sentence will be modified in the revised manuscript to make it clearer.

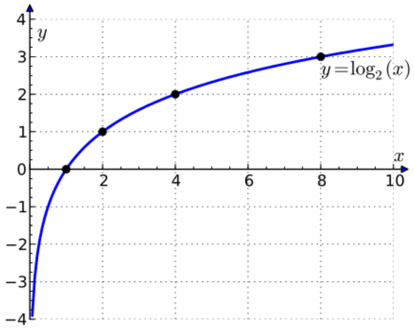
9. pg 5, ln 155: Avoid starting sentences with ”Because”. This is a general comment also through the manuscript. Would recommend re-writing this initial sentence, breaking into sections.

   We acknowledge the critiques and will account for it throughout the revised manuscript.

10. pg 6, ln 171: Can references be provided for some of these conclusions? The linkage of snow depth only to precipitation requires some basis. There are a lot of varied physical processes for snow accumulation between tundra and taiga eco-zones.

    In the revised version of the manuscript, we will provide further information on the specific characteristics of the different snow classes and refer to Sturm et al. [1995], who describe snow related characteristics for each snow class.

11. pg 8, ln 222: I don't follow the second sentence, though can be my ignorance. Consider a clearer explanation if including.

    f is a probability density function (pdf). Pdfs output only values between 0 and 1. The log_2 function returns a value less than or equal 0 for values between 0 and 1, as presented on the above plot. Subsequently, the negative of the log_2 function returns only values greater than or equal 0. We will rephrase the sentence such that less mathematical formulas are included, but will not extend the explanation, since the theory section is already quite long.

12. pg 12, Figure 3e. It is possible to rescale the number of records for each site? It is not very descriptive. Is the maximum records up above 3000?

    One station within the data set has got 3203 records. Therefore, we would like to keep Figure 3(e) as it is to represent the entire data set. However, we acknowledge that the figure can be confusing. Therefore we will include the mean and maximum in the text when the other subpanels of the figure are discussed.

13. Ordy et al had recommended the inclusion of additional explanatory variables from meteorology, such as wind or solar radiation. This study has included ERA5 daily averaged snow density data. This output from a physically based model is included without description of its generation, assessment of the quality or the relevance or applicability of this data source. Although the variable is kept as least important for the conversion model, and minimal impact on the ignorance score, it is kept in complete assessment. The manuscript should include some rationale for the

inclusion of this model output, and why it was chosen. Are the assumptions in producing the snow density relevant? How does this data perform compared to available measurements?

Wind and solar radiation are not available in real time, but only through reanalysis (e.g. ERA5). Since the model is meant to be close to operations, we only want to include variables that are available in real-time. Snow density of ERA5 was included as a test and the authors are pleased that this variable shows the lowest impact. After this result, no further variables from reanalysis were tested. We will give more information in the section 3.2 of explanatory variable and a small discussion in section 4.1.4 of input variable selection and mention that is will be excluded for operational use.

14. Pg 15, Table 2, To clarify, on pg 4, ln 97 it is mentioned that modifying the order of the input data is recommended. Is that done in this study (Shuffling data before each epoch)?

    Yes, this is part of the tested features in our study, as presented in Table 3, with the results in Table 4. However, this should become clearer when Section 3 is rearranged according to a comment by reviewer #1: 3.1 Data availability; 3.2 Input uncertainty (we would like to break up the section. In 3.2 we explain only how the input uncertainty is modelled); 3.3 Explanatory variables; 3.4 Tested characteristics. We will combine Table 2 and 3. We will also include the treatment of the input uncertainty within the MLP, the input variable selection and the determination of number of epochs and number of hidden neurons, because these are all characteristics being tested in the study and being presented in the new combined table. In 3.4 Tested characteristics, we will work with subsections to get a better structure in section 4.1.

15. Pg 17, Figure 5. MBE should be introduced before used. The use of error metrics is not entirely clear (MBE, MAE, RMSE) compared with clearly rational and description for other scores. Clearer rationale and explanation would help. Is there a reason the RMSE is shown compared to the objective function MSE? RMSE can be a more comparable error metric for SWE, but this is not explained.

    We will introduce the metrics MAE, RMSE and MBE briefly in the Section 2.4 Model evaluation. RMSE is used here because it can be compared to MAE. RMSE penalizes large residuals compared to MAE. Also, RMSE has the same units as SWE and MAE whereas MSE does not and therefore, does not have any physical meaning.

16. Pg 20, line 449: This appears a notable and relevant finding (what explanatory variables are ultimately useful) that can be better articulated in study findings.

    We will emphasize in the revised manuscript that five out of the six most important variables are coherent with the variable selection in Odry et al.

[2020]. However, we would also like to emphasis that the order of variables with scores lying close together can change, since the parameters are initialized randomly. Therefore, we can give a rough estimate of which variables are the most useful, but not an ultimate one. We will add this information to the revised version of the manuscript.

17. Pg 27, Figure 13: Consistency would be useful for interpretation between RB and MBE. Can see in the following paragraph why RB is substituted for MBE, but would like to see this graph included.

    Unfortunately, we are not entirely sure we understand this comment correctly. We disagree with the suggestion of presenting the MBE over different snow classes. Different snow classes show different magnitudes of SWE, as presented in Figure 2. Therefore, showing the MBE disables a comparison between snow classes, as the MBE will necessarily be proportional to the magnitude of SWE for a given class. The MBE can only serve for a comparison between the models for each snow class individually.

18. Pg 27, line 555: The conclusions drawn in this section appear to have a relatively weak causal or testable links. Ranging from regression model structure, to physical processes to reference model performance, several comments seem quite speculative. For example, the tundra region has poor performance, but would be subject to may be similar topographic controls of the prairies. It would seem better to reflect on what information can truly be derived from these results, or at least address that there are many contributing factors that are not represented by this method.

    We acknowledge the critiques and will delete or rephrase the speculative comments and focus on the actual output of the figure when discussing it in the text.

19. Pg 28: line 570: This opening sentence for the Conclusions section should be more descriptive and engaging in the content of the study

    In accordance with comment 1 and 4, we will also change the first paragraph of the conclusion accordingly and take the reviewers suggestion into account.

20. Pg 29, ln 591: What is the additional geophysical information beyond snow class from Sturm et al.? If this refers the discretization by elevation class, this should be elaborated on in the rest of the document to include in conclusions.

    The geophysical information is added by distributing the model into different snow classes. No further information was added. The reviewer's comment showed us that the formulation is misleading and we will delete it in the revised manuscript.

21. Pg 29, ln 596: These statements are quite generalized, and should be refined. What variables should be added and what information content

due they bring? What information is missing that could be provided by other sources and why are they not now included?

We would like to first test temperature and precipitation variables with longer time ranges. Further, we only want to look for variables which are available in real time or site specific. For instance, one could try topological variables, like slope or aspect. Also, one could test the model with more precise meteorological data on a limited number of sites. We will include these thoughts into the revised manuscript.

22. Pg 29, general: What are the limitations of the study? What is it's applicability?

    We will include some thoughts about the limitations, applicability and transferability in the conclusion. Regarding the applicability and transferability, the used data set contains records for all snow classes (except of the ice snow class being to small for a proper analysis) which shows the diversity of snow patterns within the study. Therefore, the model structure is expected to be applicable to other areas in the world. However, new training is advisable. Regarding limitations, both models show bad simulation results for high values, because the amount of training data is low. With knowing this weakness of the model, this problem could easily be bypassed in an operational context with some sort of threshold above which only the background is used, for example. Furthermore, the incorporation of climate change needs to be done manually or by taking data only from recent years. The amount of data needed to train the model properly cannot be determined universally. It depends on the variability of the data set. For instance if an area shows many different snow patterns, more data is needed to get a satisfactory result. This also changes the number of epochs and number of neurons needed to get the best result. We always advice to check the model by a validation data set which is already required in many ANN libraries in Python.

References

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning - Chapter 5 - 8*. MIT Press, 2016. http://www.deeplearningbook.org.

J. Odry, M. A. Boucher, P. Cantet, S. Lachance-Cloutier, R. Turcotte, and P. Y. St-Louis. Using artificial neural networks to estimate snow water equivalent from snow depth. *Canadian Water Resources Journal / Revue canadienne des ressources hydriques*, 0(0):1–17, 2020. doi: 10.1080/07011784.2020.1796817. URL https://doi.org/10.1080/07011784.2020.1796817.

T. H. Painter, D. F. Berisford, J. W. Boardman, K. J. Bormann, J. S. Deems, F. Gehrke, A. Hedrick, M. Joyce, R. Laidlaw, D. Marks, C. Mattmann,

B. McGurk, P. Ramirez, M. Richardson, S. M. Skiles, F. C. Seidel, and A. Winstral. The airborne snow observatory: Fusion of scanning lidar, imaging spectrometer, and physically-based modeling for mapping snow water equivalent and snow albedo. *Remote Sensing of Environment*, 184:139 – 152, 2016. doi: 10.1016/j.rse.2016.06.018.

T. Schaul, I. Antonoglou, and D. Silver. Unit Tests for Stochastic Optimization. *International Conference on Learning Representations (ICLR)*, 2014. `https://arxiv.org/abs/1312.6055`.

M. Sturm, J. Holmgren, and G. E. Liston. A Seasonal Snow Cover Classification System for Local to Global Applications. *Journal of Climate*, 8(5):1261–1283, May 1995. doi: 10.1175/1520-0442(1995)008¡1261:ASSCCS¿2.0.CO;2.