Authors' specific response regarding revision #1 to comments by Anonymous Referee #1

March 9, 2021

Black text: Reviewer's comment

Blue text: Authors' response; The identifications of lines, figures and tables refer to the version with track changes.

1 General Comments

This manuscript describes a novel method for estimating snow water equivalent from snow depth and a variety of other variables, by using an ensemble of artificial neural networks (ANN). This type of machine learning model is becoming increasingly popular as computational power increases. Likewise, the estimation of snow water equivalent from remote sensing or other in-situ variables is in constant development due to the scarce availability of in-situ SWE measurements. The manuscript falls within two topics of current great scientific interest and within the scope of the journal. I first have to congratulate the authors for the huge amount of work that they have done in the analyses and the text. This is a follow-up study to Odry et al. (2020), who introduced the use of ANNs to estimate snow density from snow depth over Quebec, outperforming other regression models such as Jonas et al 2009 and Sturm et al 2010. Here, the authors develop and improve the initial model. As elements of novelty, the authors: (1) perform an in-depth analysis of model architecture, testing several options of model characteristics; (2) demonstrate that training a model for different snow classes improves model performance; and (3) use SWE as target variable instead of snow density, which also improves performance. In general, the text is well structured and well written. The introduction and literature review are lengthy but are useful for a non-specialised reader to understand the theory behind multilayer perceptrons (MLPs) and the model ensemble evaluation metrics. The experimental set up is very detailed, although it lacks some clarity in its structure. The results are extensive and support the conclusions reached by the authors. However, the authors should be more convincing about the scientific progress that their analyses provide, as compared to Odry et al. 2020.

Moreover, the results are generally very descriptive but some more discussion (or a discussion section) lacks. The results and conclusions are focused on the model performance improvement, but there is no discussion about the limitations of these type of models (e.g. the amount of data or computation time they require). There are also quite a few specific issues to be addressed to improve the quality, clarity, and reproducibility of the study. Overall, the work presented in this study is impressive, but the authors should address the issues that I outline here before I can recommend it for publication.

We would like to thank the reviewer for their extensive review including detailed comments and suggestions. It will strengthen the output of the study. Below we address each specific comment and explain how we incorporated them into the revised manuscript.

2 Specific Comments

1. " Lines 74-79: Here you should state more clearly what the novelty of this paper is compared to Odry et al. 2020. Before I finished reading the entire paper, I was not convinced that there would be enough novelty in this manuscript. Perhaps a good solution would be to add a few lines before this paragraph, summarising the key knowledge gaps that you are filling (testing model structures, target variables, including climate classifications,...), and why it is important to tackle these knowledge gaps. What is the aim of the paper besides just “improving a model”?

   We agree with the reviewer and restructured the paragraph including more specific information, which summarizes the results of the study. In ln. 12-15, we added a more precise summary of the study in the abstract. The same information is presented in the introduction in ln. 84-92 in a broader fashion.

   It is correct that snow depth-SWE measurement sites have got roughly 100 records on average. However, in our model each record is treated individually, as the station ID is not an input in the MLP. Therefore, all data from all sites (belonging to a specific snow class for the multiple MLP ensembles model) are used to train and validate the model. The model is not trained on 100 records, but on much more (see Table 5 of the track change version for the number of records per class). We would also like to add some precision regarding continuous time series, the length of time series and why it is not central in this study. When using time series in our model, no improvement is expected because of the nature of the time series, but because of the greater amount of data and probably more consistent measurements. To make use of a time series, one would need to use the snow depth of the previous days as an input of the MLP, which is not possible at the moment in Canada, as the Canadian snow survey includes only a few continuous time series, mainly in British

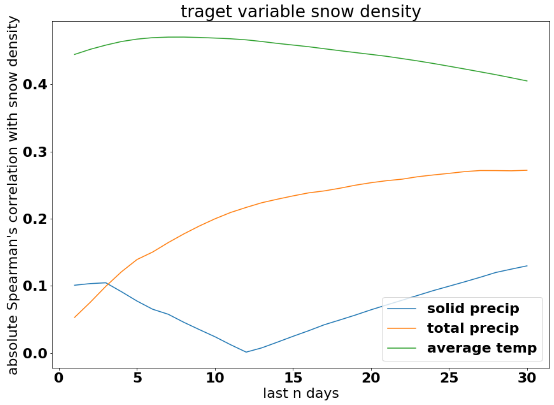
Columbia due to snow pillows measuring SWE. Elsewhere, the data is not continuous in time. Our model was developed to use only what is available in operations in Canada and therefore only makes use of data available in real-time or near-real-time. We added information about the mean and maximum of the number of records per site in the text in ln. 353-354 and in the caption of Figure 3.

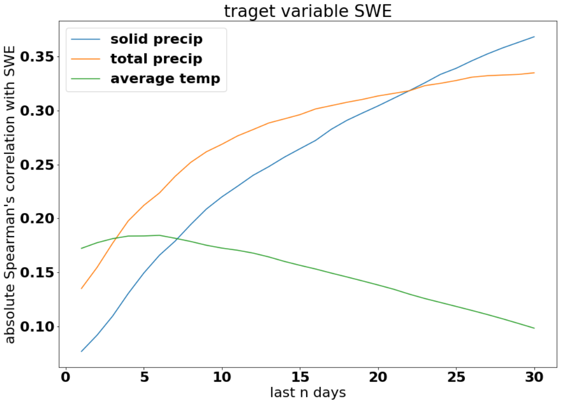
2. The explanatory variables in section 3.2 are identical to those in Odry et al. 2020, with the addition of snow density from ERA5, which turns out to be the least explanatory in Table 5. Why did not you include other variables suggested in Odry et al. 2020 such as wind and solar radiation?

   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 operational capabilities, we only want to include variables that are available in real-time. ERA5 snow density was primarily included as a test and the authors are pleased that this variable shows the lowest impact. After this result, no further variables from reanalysis where tested. We added in the section 3.2 ”Explanatory variables” in ln. 370-371 that snow density from ERA5 is included as a test. In the result section in ln. 525-526, we give the information that snow density can be excluded for operational use.

3. For variables of accumulated precipitation in the last n days, n is only tested for up to 10 days. In Table 1, three out of four variables show largest correlation for 10 days, which makes me suspect that the range tested is not big enough. Please test for a larger range. In case this becomes increasingly large, it could be that accumulated precipitation in the last n days and accumulated precipitation since the beginning of the winter are the “same explanatory variable”.

   The correlation was only calculated for a range of 1 to 10 days and no further investigation was done during the study. The two plots below show the absolute Spearman’s correlation for an extended range of 1 to 30 days. As mentioned by the reviewer, the correlation for solid and total precipitation increases with the number of days and therefore will lead to the same explanatory variable accumulated precipitation since the beginning of the winter. Note that also for snow density the Spearman’s correlation increases after 12 days and overreaches the first maximum after 27 days. However, this shows that recent solid precipitation and accumulated precipitation over the entire winter season carry different information when looking at snow density. This is also pointed out by Odry et al. [2020] their Table 4. The impact of short term and long term variables with respect to SWE need to be reexamined. We mention in section 3.2 ”Explanatory variables” in ln. 395-397 that we want to keep the information of short term variables. The above discussion is given in the conclusion in ln. 702-709.





4. The structure of Section 3 is somewhat confusing. I don't clearly understand what is a "tested characteristic" and what is not. For instance, input uncertainty and input variable selection are two "characteristics" that are tested according to Table 2 and 3, but they have their own subsection (3.3 and 3.5 respectively), while the other tested characteristics (e.g. optimization algorithm) are described in section 3.4. I suggest blending section 3.5 and 3.2 together, and to include 3.3 and 3.6 in 3.4. Combining Table 2 and Table 3 together would also help understand what is being tested and what is not. Please either restructure the section, or clearly justify the current one.The results in section 4.1 should then be restructured according to the new structure of section 3.

   We very much appreciate the suggestions, and we agree that the structure of the section can confuse the reader. We restructured section 3 as follows: 3.1 Data availability; 3.2 Explanatory variables including input uncertainty; 3.3 Tested characteristics (we combined Table 2 and 3. Table

2 now shows all tested characteristics including input uncertainty within the MLP, the input variable selection and the determination of number of epochs and number of hidden neurons) This corresponds to ln. 399-446 and Table 2.

5. Table 4: the combination (combo) of the best choice for each tested characteristic provides even better performances for most evaluation metrics. Why are all the characteristics tested one by one? Please test the combined effect of characteristics or provide a clear explanation about why this is not possible (computationally too expensive?) or not necessary.

   The characteristics have been tested one by one to measure the effect of each characteristic individually, following the Ceteris paribus principle. After that the combination of the characteristics were tested where improvement has been shown. Additionnal testing of other combinations might improve the system further. However, no large improvement is expected, as the individual tests of the other characteristics showed no improvement. This information is given in ln. 425-427. We would also like to mention that the testing is computationally expensive because it runs over several numbers of epochs (mostly 2-200).

6. Line 472-473: What part of Section 4.1 are you referring to? It is not even finished, 4.1.6 is still coming after this. I suggest to provide an extra table (or a paragraph) with the final MLP architecture set-up, to avoid having to jump section by section to gather all the final decisions.

   For less confusion, we included the section 4.1.3 ”Final setup of SMLP and MMLP” in ln. 564 and present the final setup in Table 6.

7. ”There is barely any discussion about computational cost of the final model architecture set-up, and about the number of epochs and hidden neurons. If applicable, I would like to see what the computational tradeoff of the final model choices is (e.g.choosing for XXX is about 103 times slower than YYY, though I acknowledge this is CPU dependent).”

   The discussion about computational cost for the two final models is presented in the section 4.1.3 ”Final setup of SMLP and MMLP” in ln. 566-571.

8. Line 482-483: and the number of neurons decreases too. Please explain why.

   We included in ln. 561-563 that snow classes with smaller datasets show smaller variability in the records, which can be easily represented by a simpler network with less hidden neurons, because of the lower complexity of the problem.

9. Lines 489-491: Why and what does it mean? This is the only reference to Figure 10 in the text. Please provide more information on the results shown in the figure. A discussion of the results is also necessary here, since there is no Discussion section.

We expanded the discussion on Table 7 and Figure 10 in ln. 574-582. Further, we tried to be more precise in the caption of Figure 10. We indicated which subfigures are reliability diagrams, which are rank histograms, etc. We included some discussion on the results of the rank histogram and the reliability diagram including a link to section 2.4.4 and 2.4.5, where the two evaluation metrics are introduced.

10. Table 7: How are the values obtained for MMLP? Is it the median of the ensembles for each snow class? Please specify

    In both models (SMLP and MMLP), the median is taken for MAE, RMSE and MBE. We introduced MAE, RMSE and MBE briefly in section 2.4.1 (ln.243-253), because this was asked by reviewer #2. In there, we mention that the median is taken. For clarification, when simulating the test data set for each record, the snow class is determined and the associated MLP ensemble is taken in the multiple MLP ensembles model. This returns one ensemble for one record, as in the single MLP ensemble model.

11. Figure 10: top row: How is the "median simulation" obtained for a specific bin? How is the histogram of medians built? This is highly unclear, please specify. Also,why is there a small bin on the negative side? Does the model simulate negative SWE values? Middle row: Is this a reliability diagram? It can be guessed from the text but it is not specified anywhere else. Please provide a letter for all subpanels (a-f).

    For clarity, we added letters to all sub-panels and provided more information in the caption in Figure 10. Also, we included more information about the calculation of the median simulation in the text in ln. 576. For each record, one ensemble is simulated from which we calculate the median. Further, both models simulate negative SWE values. For the single MLP ensemble the minimum of the simulation is $-36mm$ and for 0.6% of the records in the testing data set the model simulates negative SWE values. For the multiple MLP ensembles model the minimum is $-42mm$ and the ratio of negative simulation is 0.3%. This information is presented in ln. 577-579. As a side note, it is possible to get negative values for SWE, because the output layer is modelled by a linear function and therefore can output any value.

12. ”Lines 495-496: Looking at figure 11 the values seem to be rather +/-18 and +/-16? Please provide the accurate values.”

    The exact values of the box are $[-17.1mm, 18.4mm]$ and for the whisker $[-77.6mm, 73.3mm]$ for the single MLP ensemble model. The exact values of the box are $[-15.8mm, 18.0mm]$ and for the whisker $[-68.8mm, 64.6mm]$ for the multiple MLP ensembles model. This information is added in ln. 585-588.

13. Line 552: In what figure/panel do you see the better accuracy for ephemeral snow? Please add this information.

We think that the reviewer is referring to line 502. The discussion refers to Figure 12(a) and is included in ln. 594.

14. Line 507: Where is the deeper analysis of the ephemeral snow class shown? The text refers to a rank histogram for the snow class, but I don't see it. Same in line 520 referring to a rank histogram for mountain and maritime snow, where is it? Please show them.

    We apologize that the rank histograms and are not presented in the manuscript for the *ephemeral, maritime* and *mountain* snow class. We have tried to keep the manuscript as concise as possible. We included Figure 13, showing the rank histogram for SMLP and MMLP within the *ephemeral* snow class. The corresponding discussion is given in ln. 599-601. We deleted the sentence which refers to the rank histogram for the *mountain* and *maritime* snow class, since no big difference is seen between the two snow classes, which already reflects the similar skill score.

15. Figure 12: I believe the Y axis in panel (a) should be Skill Score (SS). Just as you did in Figure 13, please write "Skill Score" either on the axis label or in the caption, it is useful for the reader to be reminded what "SS" means. Same with RB, I had to read the text to find what it means (Relative Bias). Why the SMLP is evaluated on RB and MMLP on MBE-SS?

    Our initial idea was to separate the SWE error metrics (MAE, RMSE, MBE) from the CRPS and the ignorance score, as it is written in the legend. We changed the y-axis to Skill Score (SS) for panel (a),(b) and (c) in Figure 12, which is better understandable for the reader, as proposed by the reviewer. We also included ”relative bias (RB)" in the caption. Further, we would like to apologize for the mistake. In both cases the relative bias was taken. This was corrected in the revised manuscript in Figure 12.

16. Figure 14: Given the large amount of data, consider adding color to the scatter plots based on the density of points.

    Figure 15 was updated to a color coded scatter plot based on the density of points.

17. Line 562: How many are "numerous"? It seems as if they are < 10? In that case, the MLP probably can't be trained for those high values of SWE, while the regression model is continuous also for high values. Please discuss the lack of training data for these high values.

    In the testing data set there are 17 SWE measurement above $2500mm$. The training data set includes 18 SWE measurements above $2500mm$. Little training data in the higher range of SWE disables the MLP to estimate them correctly, because during training the model, it is adjusted such that the MSE over all data points is minimised. Therefore the model focuses on areas where the density of data point is the highest. This information is included in ln. 654-658.

18. TITLE: The title of the manuscript is still too similar to Odry et al. 2020. Think of a title that would directly show the improvement/novelty with respect to Odry et al.2020. (e.g. use “testing ANN architectures”, “snow class”, “climatological variables” or“multiple ensembles”)

    The title is changed to ”Investigating ANN architectures and training to estimate SWE directly from snow depth”.

19. CAPTIONS: Throughout the manuscript, Figure and Table captions are just one line long. Please extend them. Make sure everything (acronyms, lines, points, etc.) that is shown in the Figure or Table is described on the legend, labels, or in the caption.

    We believe we have already partially addressed this comment with our previous answers. More information is provided in the captions of Figures 3, 5, 6, 7, 8, 9, 10 and 12.

20. ”DISCUSSION: Following many of my comments, provide discussion of results within the results section, or provide a separate discussion section before conclusions.I especially miss a discussion on limitations of the model, applicability and transferability. How much data does the model need in order to be properly trained? You could for instance re-do the analyses but using only a random 10”

    We would like to avoid a discussion section, because we think it is more convenient for the reader when the results are discussed when they are presented. We extended our discussions in the result section in correspondence with the above comments. Furthermore, some thoughts about limitations, applicability and transferability are included in the conclusion. Regarding the latter two, the used data set has got data records of all snow classes (except the ice snow class, for which we have too little records 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. This information is included in ln. 696-699. Regarding limitation, both models show bad simulation results for high values, because the amount of training data is low. Furthermore, the model is not predictive and especially cannot account for the effect of climate change. The amount of data needed to train the model properly cannot be answered 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 advise to check the model by a validation data set, which is already required in many ANN libraries in Python. This information is included in ln. 713-720.

3 Technical corrections

- Line 75: Please replace "verify" for "test", otherwise it seems that the hypotheses have been customised based on your results.

  This is changed in ln. 86.

- Line 192: I'm guessing this is a typing error, otherwise why is DOYobs = 0 on 1st January? Shouldn't it be 123 provided 1st September is DOY = zero?

  It is not a typing error. DOY_{obs} is 0 in 1st of January and takes values from -122 till 243. This is consistent with the model proposed by Sturm et al. [2010]. However, we acknowledge our explanation was much too brief and therefore, it is explained more precisely in ln. 213-214.

- Line 457: Figure 8a-e Figure 8: panel g should be f, and h should be g. Also, I suggest making a 4x2 (or 2x4) panel figure, with the only empty panel filled with the legend. It would reduce white space.

  We welcome the suggestion of the reviewer and changed the layout and corrected the letter identifications of the subplots in Figure 8.

- Line 526: "Second, [...]" but where is first? Rephrase accordingly.

  We rephrased this in ln. 618.

  item Line 588: Remove "the remainder of".

  We removed "the remainder of" in ln. 689.

References

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`.

M. Sturm, B. Taras, G. E. Liston, C. Derksen, T. Jonas, and J. Lea. Estimating snow water equivalent using snow depth data and climate classes. *Journal of Hydrometeorology*, 11(6):1380–1394, 2010. doi: 10.1175/2010JHM1202.1.

Authors' specific response regarding revision #1 to comments by Anonymous Referee #2

March 9, 2021

Black text: Reviewer's comment

Blue text: Authors' response; The identification of lines, figures and tables refer to the version with track changes.

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 tried to take their comments into account, which we believe strengthened the output of the study. In the following, we address each specific comment and explain how we incorporated 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 summarized the outcomes of the study more precisely. Specifically, we first focus on using SWE instead of density as the target variable. Second, testing several options of ANN structural characteristics (e.g. optimization algorithm, activation function, parameter initialization, increasing the number of parameters) improves estimates of SWE. Third, including input uncertainty on snow depth improves the model's performance and dividing the area into snow classes. Fourth, using an individual ANN model for each of them gives a greater representation of the geophysical diversity of snow. This information is included in ln. 12-18.

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

   Physics-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, physics-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. These thoughts were added in ln. 54-59. 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 included this article into the introduction in ln. 60-62.

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 tried to be more specific with the outcomes of the study in ln. 87-92.

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 refined the explanations in the revised version of the manuscript accordingly in ln. 92-95.

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

   This sentence was deleted and the information was reformulated in accordance with the previous comment, referring to ln. 92-95.

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” was included in ln. 111, 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.

   The statement is 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 is tied to the previous in ln. 149.

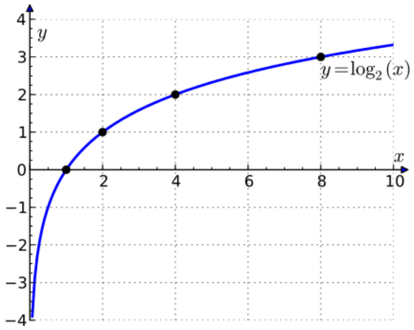
9. pg 5, ln 155: Avoid starting sentences with ”Because”. This is a general comment also through the manuscript. Would recommend re-writing this initial sentence, breaking into sections.

   We acknowledge the critiques and accounted for it throughout the revised manuscript in ln. 144-145, 173-176, 184-185, 324-325, 332-334, 358-359.

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.

    We provided 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 in ln.193-195.

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 rephrased the sentence such that less mathematical formulas are included in ln. 257-256.

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 site 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 included the mean and maximum of the number of records per site in the text in ln 353-355 and in the caption of Figure 3.

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 assump-

tions 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 operation, 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 added in the section 3.2 ”Explanatory variables” in ln. 370-371 that snow density from ERA5 is included as a test. In the result section in ln. 525-526, we give the information that snow density can 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. The Section 3 is rearranged according to a comment by reviewer #1 to the following: 3.1 Data availability; 3.2 Explanatory variables including input uncertainty; 3.3 Tested characteristics (we combined Table 2 and 3. Table 2 shows now all tested characteristics including input uncertainty within the MLP, the input variable selection and the determination of number of epochs and number of hidden neurons) This corresponds to ln. 399-446. Now, Table 2 shows the *Reference* setup and all ”Options” being tested throughout the study.

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 introduced the metrics MAE, RMSE and MBE briefly in the Section 2.4.1 in ln. 243-253. 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 emphasized in ln. 513-515 that five out of the six most important variables are coherent with the variable selection in Odry et al. [2020]. However, we also emphasized 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.

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 if 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 deleted or rephrased the speculative comments with focus on the actual output of the Figure 14 when discussing it in the text. The changes are done in ln. 647-650.

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 also changed the first paragraph of the conclusion in ln. 686-688 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 deleted it in ln. 692.

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?

    After proposing SWE as the new target variable in this study, short term and long term variables regarding precipitation with respect to SWE need to be analyzed. This also indicates Table 1. Furthermore, we only want to look for variables which are available in real time or site specific. E.g. topological variables like the slope and aspect of measurement site can be used. These thought and further discussion are included in the manuscript in ln. 702-712.

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

    Some thoughts about limitations, applicability and transferability are included in the conclusion. Regarding the latter two, the dataset used in this study has data records of all snow classes (except the ice snow class, for which we have too little records 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. This information is included in ln. 696-699. Regarding limitation, both models show bad simulation results for high values, because the amount of training data is low. Furthermore, the model is not predictive and especially cannot account for the effect of climate change. The amount of data needed to train the model properly cannot be answered 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 advise to check the model by a validation data set which is already required in many ANN libraries in Python. This information is included in ln. 713-720.

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.