Authors' response to interactive comments by Anonymous Referee #1

January 26, 2021

Black text: Reviewer's comment

Blue text: Authors' response

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 will address each specific comment and explain how we will incorporate 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 will restructure the paragraph to include more specific information summarizing the results of the study. After presenting the key knowledge gaps, we will emphasize the main outputs of the study more precisely. The focus will be on using SWE instead of density as the target variable, Testing several options of ANN structural characteristics, including input uncertainty on snow depth and dividing the area into snow classes by using an individual ANN model for each of them to give a greater representation of the geophysical diversity of snow.

2. As estimated from line 309 and Figure 3(e), snow depth-SWE measurement sites have on average 100 records. This means that the model is mostly developed over independent and sparse measurements, rather than with time series. The influence of this is not discussed. Would MLPs benefit from training with long term time series? What would be the influence of including back-propagation loops in that case? Please discuss it. Testing it for a long-term dataset such as the SNOTEL dataset over the United Stated would be worth, but I acknowledge this would require an entire follow-up analysis.

   It is correct that snow depth-SWE measurement sites have got 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 are used to train and validate the model. The model is not trained on 100 records, but on much more (see Table 6 of the manuscript for the number of records per class). We will include the above clarification into the revised manuscript. Regarding continuous time series, the length of time series and why it is not central in this study, we would also like to add some precision. When using time series, 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.

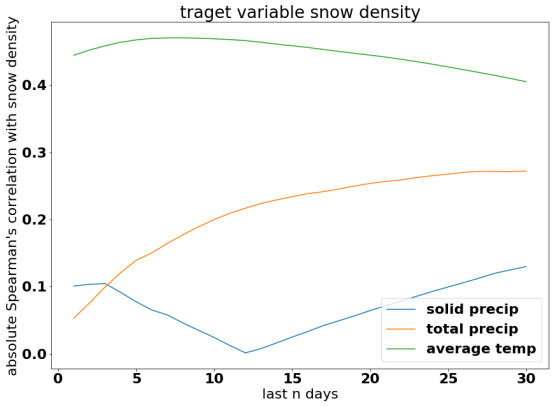
3. The explanatory variables in section 3.2 are identical to those in Odry et al. 2020, with the addition of snow density from ERA5, which turns out to be the least explanatory in Table 5. Why did not you include other variables suggested in Odry et al. 2020 such as wind and solar radiation?

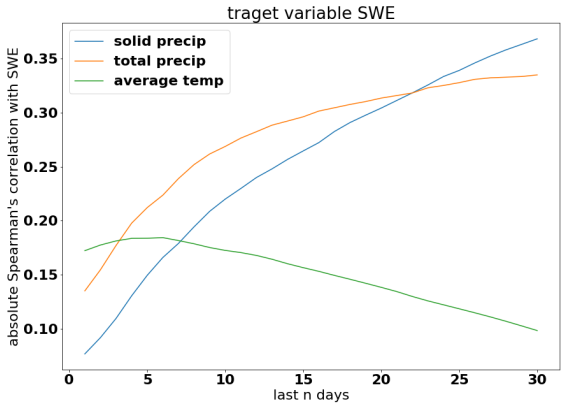
   Wind and solar radiation are not available in real time, but only through reanalysis (e.g. ERA5). Since the model is meant to be close to operational capabilities, we only want to include variables that are available in real-time. ERA5 snow density was primarily included as a test and the authors are pleased that this variable shows the lowest impact. After this result, no further variables from reanalysis where tested. We will provide more information on this in section 3.2 and a small discussion in section 4.1.4, mentioning that snow density will be excluded from operational use.

4. 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 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 does increase above 12 days and overreaches the first maximum at 3 days. We will mention this in the manuscript, but we would prefer to keep the input

variables as they are right now, as including input variables accounting for a longer time correlation would basically require recomputing everything from the start. This would unfortunately be very difficult for us at the moment, for time constraints reasons. However, the point of the reviewer is very valuable and will be considered in future studies.





5. 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 would like to propose a different structure: 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 take the suggestion of the reviewer and combine Table 2 and 3. Furthermore, we will 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.

6. 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. Further 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. We would like to mention that the testing is computationally expensive because it runs over several numbers of epochs (2-200).

7. 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.

   Two different models are built. The first uses one MLP ensemble and is applied to the entire area of Canada. This model is finalized in section 4.1.5. Section 4.1.6 finalizes the model using multiple MLP ensembles, one for each snow class. As the structure of Section 3 will be changed,

we will also change the structure here (Section 4.1). We will also provide a table providing the final set up of both models.

8. "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 trade-off 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)."

   We will provide the computational cost for the two final models (single MLP and multiple MLP) for the training on the training data set and the simulation of the testing data set.

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

   Snow classes with larger number of data points show higher variability in the records. This can be better represented by a network with more hidden neurons because complexity is increased. This information will be added to the revised version of the manuscript.

10. 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 can expand the discussion on Table 7. About Figure 10, we will clarify the caption, as we will do for all figures. We will clearly indicate which subfigures are reliability diagrams, which are rank histograms, etc. However, we are not sure what the reviewer means by "Why and what does it mean". The rank histogram is more reliable, because it is flatter with less outliers for the multiple MLP ensembles, presented by the first and last bar of the histogram. Furthermore, the reliability diagram shows a more reliable forecast when the points are closer to the identity line, which is the case for the multiple MLP ensembles. Both information are presented in section 2.4.3 and 2.4.4 when the evaluation metrics are introduced. We will provide a small explanation in the manuscript and refer to the sections 2.4.3. and 2.4.4. We hope that this will satisfy the reviewer's concerns.

11. 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 will introduce MAE, RMSE and MBE briefly in section 2.4, because this was asked by reviewer #2. In there, we will 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.

12. 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 will add letters to all sub-panels and provide more information in the caption. Also, we will provide more information about the calculation of the median simulation in the caption. 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 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 will be included in the revised manuscript, when Figure 10 is discussed in the text. 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.

13. ”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. These values will be included in the revised version of the manuscript.

14. 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 will be included in the manuscript.

15. 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 histogram and reliability diagram are not presented in the manuscript for each snow class. We have tried to keep the manuscript as concise as possible. We would prefer not to include this large figure in the manuscript itself, and we would prefer to provide it as an additional material.

16. 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. Skill Score (SS) will be written on the y-axis for panel (a),(b) and (c), which is better understandable for the reader, as proposed by the reviewer. We will also include Relative Bias in either the legend or caption. Further, we would like to apologize for the mistake. In both cases the relative bias was taken. This will be corrected in the revised manuscript.

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

    We thank you for the suggestion, we will present a color coded figure in the revised manuscript.

18. 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. We will provide a small discussion of this aspect in the revised manuscript.

19. 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")

    We suggest changing the title to "Investigating ANN architectures and training to estimate SWE directly from snow depth".

20. 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. We will provide more information for figures where it is needed in the revised manuscript.

21. "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 will extend our discussions in the result section. Furthermore, we will include some thoughts about limitations, applicability and transferability 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. Regarding limitation, both models show bad simulation results for high values, because the amount of training data is low. Further, 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 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.

3 Technical corrections

- 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 will be explained more precisely in the revised manuscript.

- All other technical correction comments are clear and will be included in the revised manuscript.

References

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.