Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2020-566-RC1, 2020

Hydrology and Earth System Sciences
Discussions

Interactive comment on "Using an ensemble of artificial neural networks to convert snow depth to snow water equivalent over Canada" *by* Konstantin Franz Fotios Ntokas et al.

Anonymous Referee #1

Received and published: 23 December 2020

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.

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"?

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.

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 didn't you include other variables suggested in Odry et al. 2020 such as wind and solar radiation?

4. For variables of accumulated precipitation in the last n days, n is only tested for 1 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".

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.

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.

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.

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).

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

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.

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

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).

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

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

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.

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?

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

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.

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")

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.

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

Technical corrections:

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

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?

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.

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

Line 588: Remove "the remainder of".

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2020-566, 2020.