Earth Syst. Sci. Data Discuss., referee comment RC1
https://doi.org/10.5194/essd-2022-63-RC1, 2022

Comment on essd-2022-63

Baptiste Vandecrux (Referee)

Referee comment on "A long-term daily gridded snow depth dataset for the Northern Hemisphere from 1980 to 2019 based on machine learning" by Yanxing Hu et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2022-63-RC1, 2022

Review of "A long-term daily gridded snow depth dataset for the Northern Hemisphere from 1980 to 2019 based on machine learning" by Hu et al.

Baptiste Vandecrux (bav@geus.dk)

General comment:

This article uses multiple gridded snow depth products and combine them with a random forest (RF) algorithm trained on more than 30 000 in situ observations of snow depth. The study covers an interesting topic as snow depth is critical for many climatic and ecological processes. The use of machine learning (ML) algorithms is also interesting for the combination of data from multiple sources. The study has therefore a clear potential as the dataset could benefit to many other research groups. However, I have major concerns on the following topic:

- The novelty of the work: It seems that the same authors have presented a similar study in Hu et al. (2021). It is unclear what are the differences between this previous study and the current work. I am unsure that adding one or two inputs to the same framework justifies a new publication. However, I also identify several flaws in the methods (see comments below) that, if fixed, could justify that the new product has improved enough since Hu et al. (2021).
- There is a confusion between snow depth and snow water equivalent (SWE) in the studies cited in the introductions. The two quantities are not interchangeable, and the authors should justify, with the appropriate studies, why snow depth is a variable relevant to monitor and difficult to observe on large scale. Work on SWE can naturally be reviewed as a related, but distinct, field of research.
- The choice and training of the RF algorithm: The RF algorithms are known to be very good at (over)fitting the training data and to perform poorly outside of their training set.

There is currently nothing in the method that ensure that the training set is representative of the conditions in which the RF algorithm is being used. It is now standard procedure to de-cluster the training data to make sure that the training data is not dominated by one specific type of sample.
- Overfitting: Nothing is said about how the hyper-parameters of the RF algorithm are being set and about any measures to prevent over-fitting. For example, neural networks can be trained with noise added to the training data to prevent overfitting. Maybe something similar exists for RF regressions. Additionally, the evaluation of the of the fused dataset using randomly selected samples cannot evaluate properly the output of overfitted algorithm because the random subset will have the same distribution, and therefore the same structure (clustering) as the rest of the training set. Only a carefully set-up spatial cross-validation can make sur that a ML algorithm works appropriately in all the different areas where it is eventually applied.
- The trend analysis has some major methodological flaws that will need to be addressed.

Considering these issues, I recommend a major revision of the paper, unless the editor considers that the necessary reshaping would deserve a new submission.

Specific comments:
l.48 "mass" of what? Please split the sentence in two.
l.50-53: The second half of this paragraph states that "knowledge on snow depth and its trends are lacking", that there are "limited surface observations" and that remote sensing methods are "inadequate". This is not exactly the current state of research in snow depth mapping because this same study builds on numerous gridded snow depth products and thousands of in situ observations. Please re-frame this paragraph and acknowledge properly the previous work.
l. 61-63: Please give references for each of these products.
l.85: "conventional" do you mean "convolutional"?
l. 79-90: Consider discussing this additional reference:
Shao, D., Li, H., Wang, J., Hao, X., Che, T., and Ji, W.: Reconstruction of a daily gridded snow water equivalent product for the land region above 45° N based on a ridge regression machine learning approach, Earth Syst. Sci. Data, 14, 795–809, https://doi.org/10.5194/essd-14-795-2022, 2022.
l.90: Hu et al. (2021). This study seems very similar to what is presented here. Please describe the study in further detail and make explicit how the presented study builds on top the previous one. What were the limitations of the previous study and what are the novelty in this new one?
l.94: Mudryk et al. (2015) compared snow water equivalent (SWE), not snow depth.
l.93: "more than 50%" in SWE, not snow depth
l.95: Mortimer et al. (2020) evaluate SWE products, not snow depth.
l.97: "Globsnow snow depth", it was the SWE that was evaluated there
l.98: "Previous assessments" which studies do you refer to?
l.107: Snauffer et al. (2018) should be added and discussed in line 79-90 where different ML algorithms are being used.
l.107-115 are partly redundant with l.79-90. They should be moved there and merged into a paragraph dedicated to ML algorithms used for snow depth retrieval.
l.115-117: Are the methods, and therefore produced data, the same as in Hu et al. (2021)?
If yes, then I think it raises the issue of the novelty of the study.
If not, then a paragraph in the intro should be dedicated to the limitations of Hu et al. (2021) and how the present study builds further and presents an improved product compared to Hu et al. (2021).

l.144-145 “In these two…” This sentence is unclear. Is the snow depth always set to 5 cm when being detected? How are deeper snowpack considered?
l.147-148: “The accuracy…” Give reference
l.166: remove “,” between “study” and “attempted”
l. 166-167: “Venäläinen et al., (2021)” This reference was not discussed in the intro when introducing the ML algorithms in snow depth retrieval.
l.181: Please give a reference for this dataset.
l.185: Please give a reference for this dataset.
l.190: Please give a reference for this dataset.
l.196: Please give a reference for this dataset.
l.200-205: These data are very important as they are the most objective way to evaluate your fused dataset. Please show on a map that they are located across a wide range of geographical locations and cover different land category for which you fit different RF models.
l.219-223: This is an insufficient level of detail for the core of your method. The documentation should be sufficient to reproduce your product. The fitting procedure and the hyperparameter selection should also be detailed to show how you avoid overfitting and to make the RF able to predict outside of its training set.

ML algorithms are very sensitive to the training data and to any imbalance therein. The training data should be de-clustered: it should be made sure that the observed snow depth covers the whole spectrum of retrieved snow depth and are located in all the elevations and all the land categories that are used as input to the algorithm. The de-clustering could be done by assigning weights to observations and or by duplicating observations from under-represented subsets.

Since your objective is to use the fused dataset for spatio-temporal analysis, a spatial or temporal cross validation should be conducted to investigate the robustness of your algorithm. This could be done by iteratively removing different regions or different years from the training set and using these removed samples for evaluation. Of course the final product should use as many samples as possible, but the evaluation of the RF algorithm is currently insufficient to build trust in its output.

L.269-275: This should be moved in paragraph 3.1. Or even in the description of the input snow depth dataset further up. Please quantify these data gaps.
L. 271 and 272: Replace “missing” by “gaps”
l. 278: “was properly…” replace with “projection was set to”
l.278: “spatio” replace with “spatial”
Section 4.1: The first paragraph about temporal availability and the last paragraph about file format could be moved as a new subsection 3.3 as it is not properly a result. it is about data availability and format.
l.289-291: In the training set of snow depth observations, many samples are redundant (f.e. daily snow courses will have similar values from one day to the next). Consequently, randomly extracting samples from the observation dataset will leave just as much information in the training set. The RF algorithm will then be very good at (over)fitting the training set and producing outstanding results on the test set. For a fair evaluation of all products, some observations should be left out from the RF training. Preferably this left out data should be representative of various geographical and natural settings to evaluate the product in different conditions. I thought that the data presented in lines 200-205 would serve that purpose?

l. 293: "..in situ observations." Add a reference to Figure 1.
Figure 1: Are these statistics applying to the same samples? Can you give their number? I understood that the original snow depth products have different spatial coverage and are sometimes missing data. Are these evaluation samples have data available for all products?
l.299-306 and Figure 2: I suggest that you present a mosaic of scatter plots with the original snow depth products it will illustrate your statement line 302-306. Please also be quantitative. What is "not very accurate" l. 305?
Figure 2: Is there any point above 250cm? You could narrow the axis' limits.
Figure 3: I am surprised by the little amount of observations in the Himalayas. Isn't there any snow depth measurements available there?
Table 2: Please present the number for transparency. NaN means "not a number". A bad number is still a number.
l. 338-339: "Compared with the original…" Please present a mosaic of scatter plots at the 7 sites and for the 6 products involved. This will illustrate properly this statement.
Figure 4: Please make these plots fit on one page.
Section 4.4: Are you again comparing the training set? If yes, then it should be moved just after the section 4.2. Please refer to Figure 5 early in this paragraph and please guide the reader to which panel each statement is related.
l. 371:. "BIAS" is a word, not an acronym. It should be lower case in all the manuscript.
Figure 5: Please add unit for bias. make bias lowercase and resize so that all panels fit on one page. The last three panels should have the same bin size as the others.
Table 3: Please provide and discuss the mean error.
Section 4.6: Consider having section 4 only for the evaluation of the dataset and a section 5 for the spatio-temporal analysis.
l. 399: "North America and Eurasia" It is unclear what is included in these two domains. Are the northern part of Africa and south America included? If they are, then the domains should be renamed in something more neutral (A & B, or west & east). Please be aware that significant snowpacks can be present in the northern Andes and in the Atlas Mountains.
I recommend making trend analysis in narrower regions (North/south America, West/east Europe, Asia with or without Himalayas...). These regions should be illustrated in Fig 6. Please present first the spatial pattern of average snow depth (without any trend) and then the trend analysis to avoid confusion.
l. 400. "There was an overall trend decrease followed by a slowly increase" From when to when? By how much? With what level of significance? Please refer to Fig 6.
Section 4.6: The first paragraph is about trends, then the following two paragraphs are about spatial distribution, then comes section 4.7 that presents trends again. Please present the spatial distribution of average snow depth before analyzing the temporal trends.
Figure 6: I am surprised that the Himalayas are not being highlighted as a deep snow area. Rearrange Fig 6 and 7 so that Fig 6 has all the maps of snow depth and Fig 7 has all the trend analysis for different seasons and for different regions.
l. 417 What is "roughly similar"? Please be specific and quantitative.
l. 420 "… significantly lower than that in winter and spring." what was the average in winter and spring then?
Figure 7: Have you investigated why spring 1984 had snow depth 30% higher than average?
Heading of section 4.7: Comparison to what?
Section 4.7: There is a confusion between the analysis of a snow depth change rate, which refers to the fitting of a linear model and the Mann-Kendall test, that is a statistical test that only tests whether the trend is monotonic or not. The Mann-Kendall test, in its original form, does not give the magnitude of the trend. It only tells if a trend is significantly positive or negative. A linear regression and the discussion of whether the fitted slopes are statistically different from zero would be here more suited. The results of this trend analysis should be discussed in more details. Are these results reasonable? Do

they match with other studies?
l. 431: I don't understand this test value. Does it apply to the hemisphere-average snow depth trend?
l.444: I don't see how this sentence is related to the rest of the paragraph.
l. 444-446: Please move to method.
l. 447: Yes, but how do you deal with it? It is not clear. Do you extrapolate or fill with a certain value the product south of 35degN? Do you only have RF models without GlobSnow south of 35degN? This should be explained clearly in the methods.
l. 447: ”In this study…” Can you elaborate? How can it be fixed in the future?
l. 451-452: “more snow survey” More data is good, better data is even better. What data would you need to make you fusion even better? Are there certain geographical areas or land type, elevation or latitude that has insufficient in situ observations? Please elaborate and please be specific and quantitative when possible.
l. 453-458.”black-box models” This is only partially true, and you raise an interesting point: how to understand and interpret the output of the ML algorithm. Tests such as the permutation feature importance (Breiman, 2001) or Shapely value (https://github.com/slundberg/shap, Strumbelj and Kononenko, 2014) would represent a valuable addition to the paper to explain which of the input snow depth data is the most important in different regions or periods.

Breiman, Leo.“Random Forests.” Machine Learning 45 (1). Springer: 5-32 (2001).

Shapley sampling values: Strumbelj, Erik, and Igor Kononenko. "Explaining prediction models and individual predictions with feature contributions." Knowledge and information systems 41.3 (2014): 647-665.

l.458: “In future study…” This sentence is not clear. Consider removing.
Section 5.2: This paragraph is not clear to me. Please provide a table that summarizes the different tests being considered and metrics that allow the comparison of the results of these different tests. It looks very similar to a down-sized spatial and/or temporal cross-validation. As mentioned earlier, this should be a key of how the algorithm is evaluated and therefore presented in greater details.
Section 5: Consider removing the Discussion section and renaming Section 4 as “Results and discussion”. The content of section 5 can be merged with existing paragraphs or if not possible, remain as seperate subsections.
l.470-471: “It the future…” That is very true, that is why the fitting procedure should be subject of extra care to avoid overfitting and to allow the RF algorithm to perform decently outside of its training set.
l.476 “Regarding the limitations…” This sentence should be moved in the previous subsection as it deals with limitations.