Review Report

Journal:	The Cryosphere
Journal's Ref.:	tc-2019-161
Title:	Real-time snow depth estimation and historical data reconstruction over China based on a random forest machine learning approach
Authors:	Jianwei Yang, Lingmei Jiang, Kari Luojus, Jinmei Pan, Juha Lemmetyinen, Matias Takala, Shengli Wu
Date:	2019-11-25
Recommendation:	Major revisions are needed

Aim of the manuscript

[1] The aim of the manuscript is (a) to test random forests in estimating snow depth in a remote sensing application and (b) to reconstruct historical snow depth in China in the period 1987–2018 (see page 5, lines 10–14).

[2] The procedure of the manuscript is presented in Figure 3.

General evaluation

[3] The aim of the manuscript (in particular the reconstruction of historical snow depth) as well as the use of a big dataset justifies its publication.

[4] The procedure followed in the manuscript is complicated, while I think that some steps are unnecessary and a more straightforward approach to the problem would achieve comparable (or even better results).

[5] Regarding the algorithmic part of the manuscript, I have some recommendations to justify certain choices of the manuscript and highlight some advantages and drawbacks of random forests (regarding most minor comments on the algorithmic part, e.g. parameters of random forests, variable importance, number of predictor variables and more, as well as why one should use random forests instead of another algorithm, please consider reading the random forests review by Tyralis et al. 2019a for more details).

[6] Furthermore, I think that the manuscript is wordy at some Sections, for instance explanation of Figures.

[7] Perhaps the reconstructed dataset could be made available online increasing the value of the manuscript.

[8] Some comments which should be discussed / addressed in the manuscript follow.

Major comments

[9] Page 8, line 10 – page 9, line 25: In general, I think that the procedure described here is complicated, while some steps may be unnecessary. In particular:

a. Random forests are fitted using 15 predictor variables in the period 2014–2015 (page 8, lines 11, 12) and then they are validated in the period 2012–2013. I do not understand the scope of this validation, considering that parameters of the algorithm have been defined earlier.

b. Random forests are used to predict snow depth in the period 2012–2018. Then a linear model is trained in the predictions of the period 2012-2018 using two predictor variables. The trained linear model is used to predict snow depth in the period 1987-2018.

In my opinion it would be more straightforward to train random forests in the period 2014-2015 using two predictor variables and then predict in the period 1987-2018. Another straightforward option would be to train a linear model in the period 2014-2015 and then predict in the period 1987-2018.

Instead, following the two-stage procedure of the manuscript, a dataset, obtained by some predictions, is used to train a new model. In these procedures uncertainties are introduced (since the dataset obtained by random forests is an approximation of the true snow depth) which are transferred to the second stage prediction. I understand that this approach gives a rich dataset to do the second stage training, however I think that the induced uncertainties are not compensated by the bigger dataset. Perhaps the manuscript could justify this approach by performing some comparisons between the one and the two-stage approaches in the period 2012-2018 or just completely use the straightforward approach.

c. Perhaps the approximation of equation (2) is suboptimal because it is based on data before 2008, while it does not include the intercept parameter. Given the big magnitude of the dataset, it is surprising that a one-parameter linear model (equation 2) would be preferable to the two-parameter model of equation (1).

Minor comments

[10] Page 2, lines 15 – 20: A proper assumption for applying random forests is stationarity. Furthermore random forests do not predict outside the range of the training sample. Therefore, the assumption of global warming is not compatible with random forests.

[11] Page 6, line 1: SSMI/S provides data in the period 2006-present according to Table 1.

[12] Page 7, lines 16 – 17: Random forests parameters are more than two.

[13] Page 7, lines 21 – 27: In general the default values (in the software implementation) of random forests parameters are good.

[14] Page 7, lines 21 – 27: In general it is suggested to use as high number of trees as computationally feasible. However, indeed the number of 500 trees is high enough in most applications.

[15] Page 7, line 27 – page 8, line 2: In general the larger the dataset, the better the predictive ability of a regression algorithm.

[16] Page 10, lines 8–12: By increasing the size of training sample one would expect that the performance of predictive algorithm would increase.

[17] Page 11, lines 4, 5: Which linear model?

[18] Page 11, lines 22–24: The comparison between random forests and the linear model is unfair considering that the latter uses less predictor variables.

[19] Page 12, lines 25–27: This procedure is not clear.

[20] Page 13, lines 3, 4: I do not understand why assigning values to the slope and intercept.

[21] Page 16, lines 8–11: It is not clear which period was used to compute variable importance.

[22] Page 16, lines 24–28: Perhaps the information added by the longitude and latitude predictor variables is already included in the remaining predictor variables (see e.g. a similar application in Tyralis et al. 2019b). In the latter study, the predictive performance was examined by comparing models with and without longitude and latitude, and the effect of coordinates was found insignificant. Perhaps, computing variable importance and predicting performance would give some explanations on the value of the remaining predictor variables and make the model less dependent on the proximity of nearby stations.

[23] Page 18, lines 1–3: In general one would expect that using more predictor variables related to the dependent variable of interest would improve the trained model. Furthermore, redundant predictor variables slightly affect random forests.

[24] Figure 6: Figures should be numbered and respective explanations should be added in the caption.

[25] Regarding the implementation of random forests, some of their disadvantages and their impact in the results of the study can be discussed (see a list of disadvantages in Tyralis et al. 2019a), e.g. they do not extrapolate outside the training range, variable importance metrics are not always reliable, as they are affected by high correlations and interactions, and more.

[26] Implemented software, software packages, libraries etc used in the study for computations

and visualizations should be cited in the references list to credit software developers.

Language

[27] Page 4, line 8: Perhaps regression instead of prediction would be more accurate.

References

Tyralis H, Papacharalampous G, Langousis A (2019a) A brief review of random forests for water scientists and practitioners and their recent history in water resources. Water 11(5):910. https://doi.org/10.3390/w11050910.

Tyralis H, Papacharalampous G, Tantanee S (2019b) How to explain and predict the shape parameter of the generalized extreme value distribution of streamflow extremes using a big dataset. Journal of Hydrology 574:628–645. https://doi.org/10.1016/j.jhydrol.2019.04.070.