Response to Reviewer Comments by Review #2 on "Real-Time Snow Depth Estimation and Historical Data Reconstruction Over China Based on a Random Forest Machine Learning Approach" by Jianwei Yang et al.

Thank you for your letter and the comments concerning our manuscript. Those comments have been very helpful for revising and improving our paper as well as providing guidance for our research. We have studied the comments carefully and have made corrections, which we hope meet with approval. We provide responses in blue below.

Review #2

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.

Recommendation: Major revisions are needed

General evaluation

1. 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).

Response 1: Other reviewers (Reviewers #3 and #4) gave similar comments. Thus, we redesigned the methodology in this study to improve this manuscript. The results demonstrate that certain predictor variables are unnecessary. There are four major revisions in the new manuscript.

1) Revision 1: scientific validation dataset

One of the major issues of the original manuscript was that the validation data are not independent temporally and spatially. Thus, in the revised manuscript, available stations were randomly divided into two roughly equal-sized parts by Matlab software (Fig. 1). The snow depth observations from training stations (342 sites) together with satellite T_B and other auxiliary data can be used to train the RF model. The measurements from validation stations (341 sites), as spatially independent data, can be applied to validate the fitted RF algorithm and the reconstructed snow depth product. Fig. 2 shows the histograms of snow depth observations from training and validation stations during the period 2012-2018. Ninety percent of the samples range from 1 cm to 25 cm. The maximum values of the snow depth extend to approximately 50 cm. However, the number of such cases is small and is therefore not evident in Fig. 2.

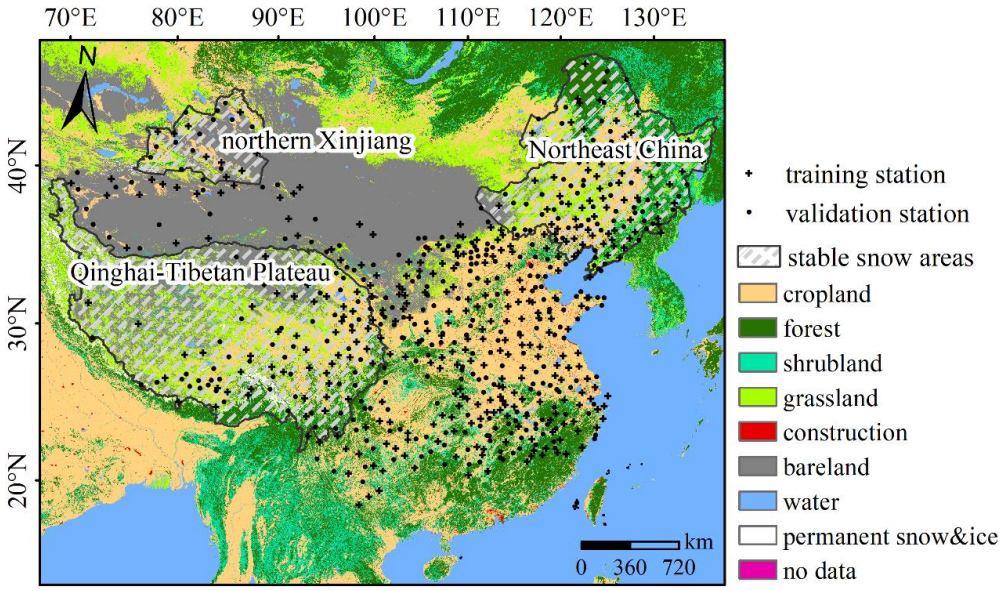

Figure 1. Spatial distribution of the weather stations and land cover types in the study area. There are three stable snow cover areas in China: Northeast China (NE), northern Xinjiang (XJ) and the Qinghai-Tibetan Plateau (QTP).

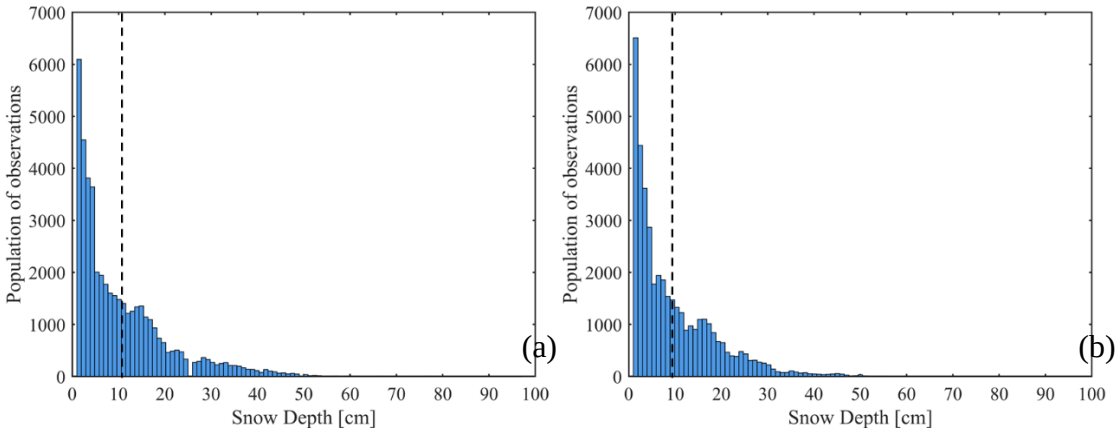

Figure 2. Histograms of snow depth observations from (a) training and (b) validation stations. The average values (black dashed lines) are equal to 10.5 cm and 9.8 cm, respectively.

2) Revision 2: four selection rules of predictor variables

The procedure described in the original manuscript is complicated. Based on the correlations between the predictor variables and the variable importance metrics (Fig. 3), we designed four schemes of predictor variables to train the RF model in the revised manuscript. The scheme one was the simplest and its predictor variables included satellite observations at 19 GHz and 37 GHz only (Table 1). The scheme four was the most complicated. We first demonstrated whether certain predictor variables are necessary and whether their inclusion affects the RF model.

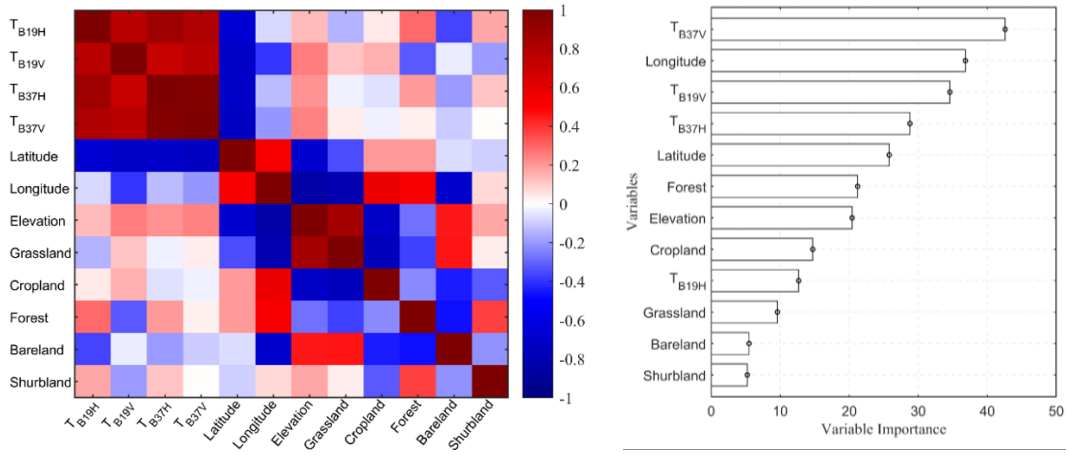

Figure 3. Correlations between the predictor variables (left) and the ranking of variable importance (right). The importance of variables, referred to as Mean Decrease Accuracy (MDA) in the RF model, is obtained by averaging the difference in out-of-bag error estimation before and after the permutation over all trees. The larger the MDA, the greater the importance of the variable is.

Table 1. A detailed description of the input predictor variables based on four selection rules of the training sample.

Name	Predictor Variables	Target	Note
RF1	T_{B19V}, T_{B37V}	snow depth	land cover types: grassland, cropland, bareland, shurbland, forest
RF2	T_{B19V}, T_{B37V}, Latitude, Longitude		
RF3	T_{B19V}, T_{B37V}, Latitude, Longitude, Elevation		
RF4	T_{B19V}, T_{B37V}, Latitude, Longitude, Elevation, Land cover fraction		

3) Revision 3: validation of the fitted RF algorithms

We conducted three tests to verify the fitted RF algorithms (Table 2). The same training samples (same algorithms) were used for the three tests but with different validation datasets. In Test1, the validation data were from out-of-bag (OOB) samples. Generally, approximately two-thirds of the samples (in-bag samples) were used to train the trees and the remaining one-third (OOB samples) were used to estimate how well the fitted RF algorithm performed. This preliminary assessment generally provides a simple way to adjust the parameters of the RF model. However, we should use the OOB errors with caution because its samples are not independent at temporal and spatial scales. In Test2, we applied temporally independent reference data during the period 2015-2018 to assess the accuracy of the temporal prediction of fitted algorithms. In Test3, a spatially independent dataset from validation stations during the period 2015-2018 was used to assess the accuracy of spatio-temporal prediction.

Fig. 4 indicates that the accuracy of RF model is greatly influenced by geographic location, elevation, and land cover fractions. However, the redundant predictor variables (if highly correlated) slightly affect the RF model. The fitted RF algorithms perform better at the

temporal scale than that at the spatial scale, with unbiased RMSEs of ~4.4 cm and ~7.3 cm, respectively.

Table 2. Summary of three tests of the fitted RF algorithms in Table 1.

Name	Test1 (OOB)		Test2 (temporal subset)		Test3 (spatio-temporal subset)	
training	training stations	2012-2014	training stations	2012-2014	training stations	2012-2014
	samples	28602	samples	28602	samples	28602
validation	training stations	2012-2014	training stations	2015-2018	validation stations	2015-2018
	samples	14301	samples	34684	samples	25879

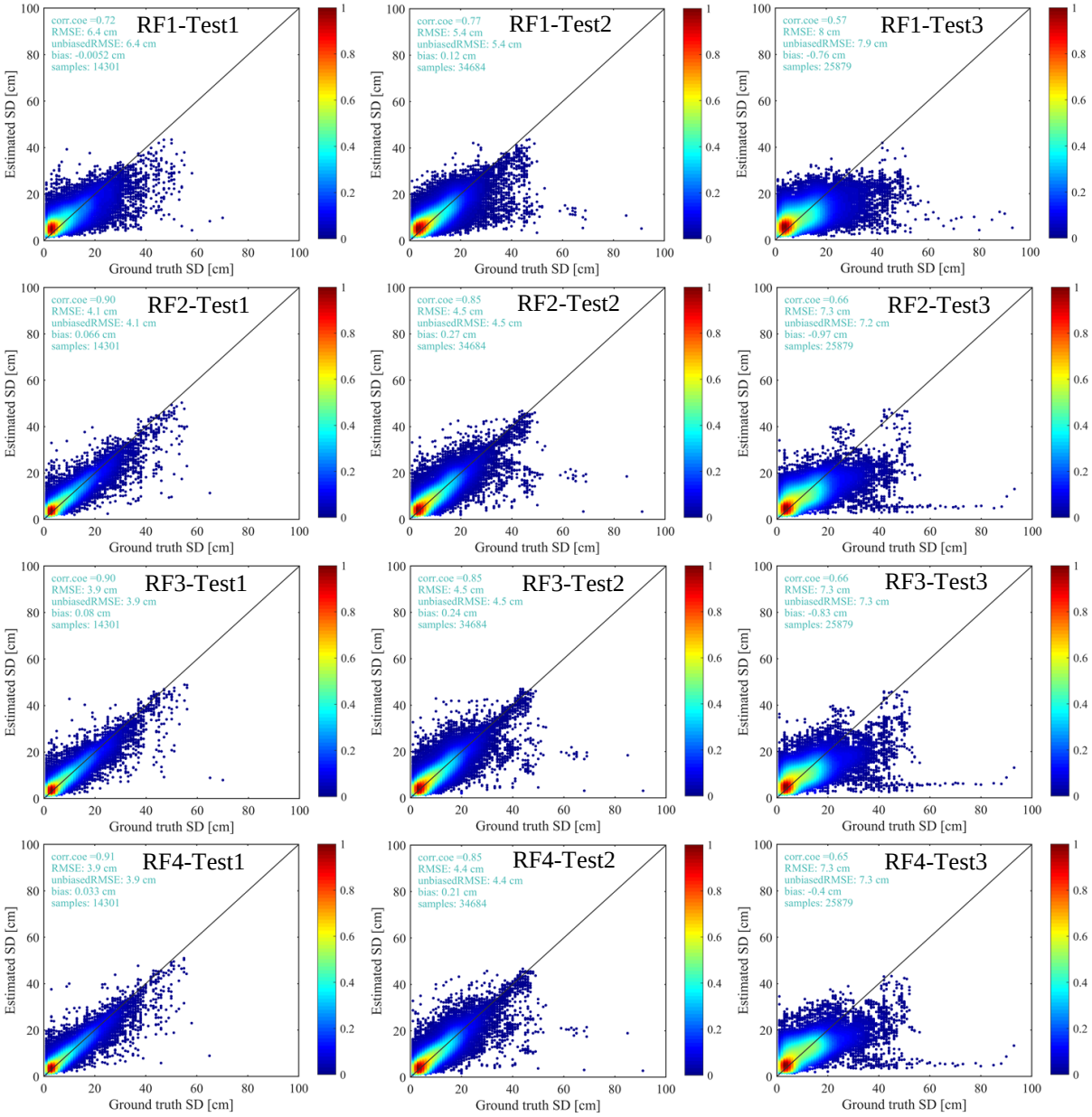

Figure 4. The color-density scatterplots of the estimated snow depth with four fitted RF algorithms and the ground truth snow depth. The four trained RF algorithms (RF1, RF2, RF3, RF4) were evaluated with three validation datasets (Test1, Test2, Test3).

4) Revision 4: validation of the reconstructed snow depth product

Finally, we directly used the fitted RF2 algorithm to retrieve a consistent 32-year daily snow depth dataset from 1987 to 2018. This product was evaluated against the independent station observations during the period 1987-2018. The mean unbiased RMSE and bias were 7.1 cm and -0.05 cm, respectively, outperforming the former snow depth dataset (8.4 cm and -1.20 cm) from the Environmental and Ecological Science Data Center for West China (WESTDC).

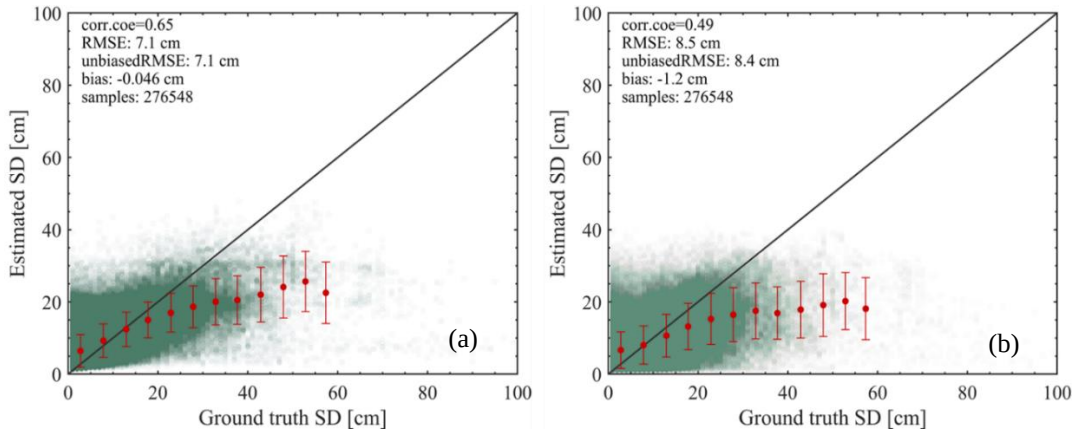

Figure 5. Scatterplots of the estimated snow depth and the ground truth observation for (a) RF and (b) WESTDC products.

To determine the interannual variability in the uncertainty, the time series of assessment indexes, including the unbiased RMSE, bias and correlation coefficient, are shown in Fig. 6. The results show that the RF estimates outperform the WESTDC product with respect to unbiased RMSE and correlation coefficient from season to season.

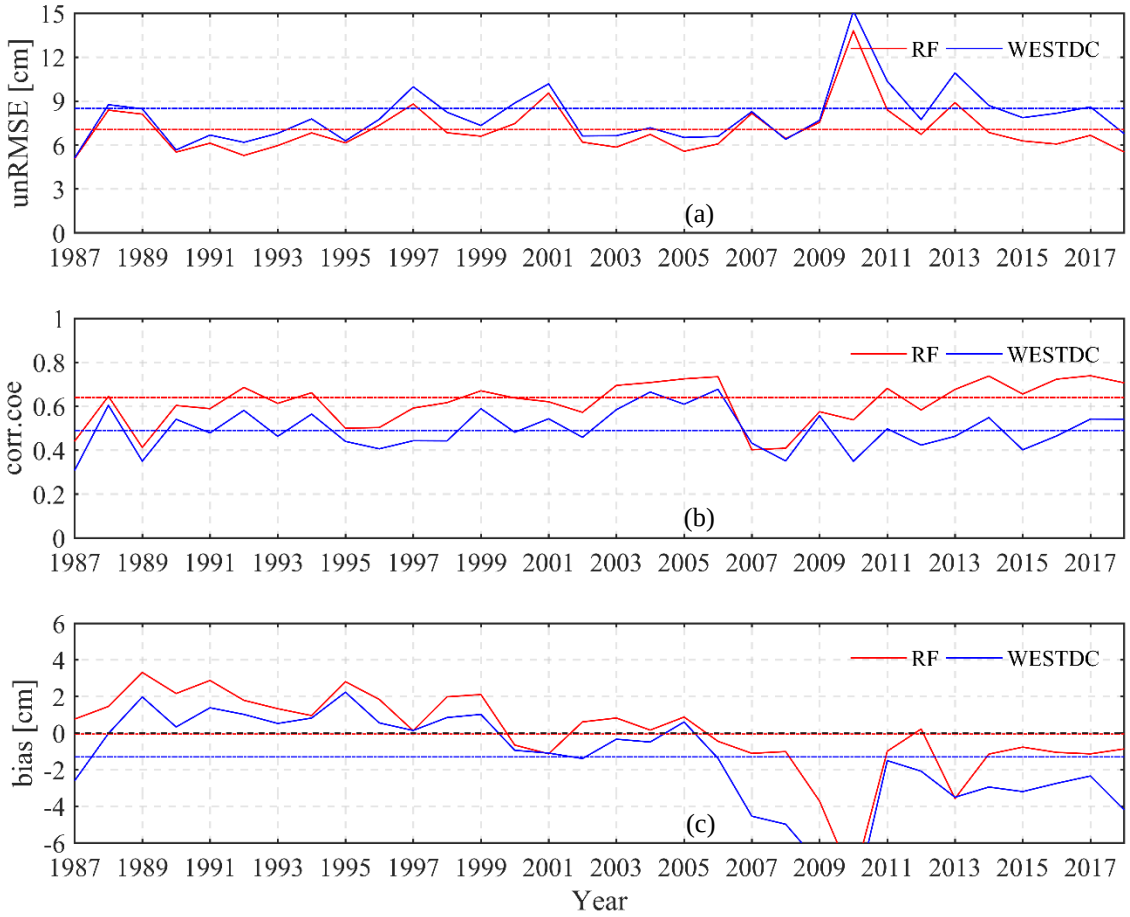

Figure 6. Time series of (a) unbiased RMSE (unRMSE), (b) correlation coefficient (corr.coe) and (c) bias for RF and WESTDC products. The colorful dashed lines represent mean values of assessment indexes.

2. 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)..

Response 2: We appreciate the reviewer's help and suggestions. We conducted a test to justify whether certain steps are necessary. Please refer to the response to "General evaluation 1" above.

We read the reference carefully. It is a good paper and was very useful for us. We have rewritten the introduction to the RF model in Section 2.2.1.

"RF is an ensemble ML algorithm proposed by Breiman in 2001. It combines several randomized decision trees and aggregates their predictions by averaging in regression (Biau and Scornet, 2016). Generally, approximately two-thirds of the samples (in-bag samples) are used to train the trees and the remaining one-third (out-of-bag samples, OOB) are used to estimate how well the fitted RF algorithm performs. Few user-defined parameters are generally required to optimize the algorithm, such as the number of trees in the ensemble (*n*tree) and the number of random variables at each node (*m*try). The *n*tree is set equal to 1000 in the present study since the gain in the predictive performance of the

algorithm would be small with the addition of more trees (Probst and Boulesteix, 2018). The default value of *m*try is determined by the number of input prediction variables, usually 1/3 for regression tasks (Biau and Scornet, 2016). The RF regression is insensitive to the quality of training samples and to overfitting due to the large number of decision trees produced by randomly selecting a subset of training samples and a subset of variables for splitting at each tree node (Maxwell et al., 2018). In addition, RF provides an assessment of the relative importance of predictor variables, which have proven to be useful for evaluating the relative contribution of input variables (Tyralis et al., 2019b). Furthermore, the RF model can rapidly trained and is easy to use. In this paper, a randomForest R package (Version 4.6-14) is used for regression (Liaw and Wiener 2002; Breiman et al. 2018)" (Page 4, Line 20-30 in the revised manuscript).

We also highlignted the drawbacks of RF model in Senction 4.1.
“The RF technique is already used to generate temporal and spatial predictions. Generally, the RF model cannot extrapolate outside the training range (Hengl et al., 2018). Fig. 6 and Table 4 indicate that the spatial predictions of fitted RF algorithms are more biased than are the temporal predictions. Thus, the transferability of a fitted RF algorithm to other areas is in question. Several studies (Prasad, Iverson & Liaw, 2006; Hengl et al., 2017; Vaysse & Lagacherie, 2015; Nussbaum et al., 2018) have proven that RF is a promising technique for spatial prediction; however, these studies aim at spatial prediction of properties that are relatively static over the observational period, e.g., soil types and soil properties.
What makes the Earth system interesting is that it is not static but dynamic (especially concerning snow parameters). Generally, snow depth increases at the beginning of winter and then decreases in spring due to melting. Moreover, snow cover has different spatial patterns in various regions, such as generally deep snow in high-latitude and high-elevation areas. In China, there are five climatological snow classes following the classification by Sturm et al. (1995). Each snow class is defined by an ensemble of snow stratigraphic characteristics, including snow density, grain size, and crystal morphology, which influences the snowpack’s microwave signature (Sturm et al., 2010). These dynamic properties of snow will lead to many cases in which the same satellite T_B corresponds to different snow depths, while the same snow depth is associated with various T_B observations, rendering the fitted RF algorithm suboptimal. Using ML techniques in combination with snow forward models (physical modeling) has the potential to overcome many limitations that have hindered a more widespread adoption of ML approaches” (Page 9, Line 20-30 in the revised manuscript).

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

Response 3: We agree with the reviewer’s opinion. We revised all the sections thoroughly to make it more precise.

4. Perhaps the reconstructed dataset could be made available online increasing the value of the manuscript.

Response 4: We agree with the reviewer's opinion. The reconstructed dataset from 1987 to 2018 is now available and we will upload the data later.

Major comments:

1. 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.

Response 1: Thank you for your comments. We have revised the manuscript. Please refer to the response to "General evaluation 1" above.

2. 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.

Response 2: Thank you for your constructive comments. In the revised manuscript, we directly used the fitted RF algorithm to retrieve a consistent 32-year daily snow depth dataset from 1987 to 2018. Please refer to the response to "General evaluation 1" above.

3. 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.

Response 3: Other reviewers gave similar useful and constructive comments. Thus, we directly used the fitted RF algorithm to retrieve a consistent 32-year daily snow depth dataset from 1987 to 2018 in the revised manuscript and omitted the pixel-based algorithm.

4. 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).

Response 4: According to reviewers' suggestions, we directly used the trained RF model to retrieve long-term snow depth product, leaving out the pixel-based algorithm. Please refer to the response to "General evaluation 1" above.

Minor comments

1. 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.

Response 1: We agree with the reviewer's opinion. We deleted this sentence.

2. Page 6, line 1: SSMIS provides data in the period 2006-present according to Table 1.

Response 2: Yes, SSMIS provides data from 2006 to the present and SSM/I from 1987 to 2008 (Table 3). We changed the sentence to the following: "The series of the Special Sensor Microwave/Imager (SSM/I) and Special Sensor Microwave Imager Sounder (SSMIS) instruments has provided continuous T_B measurements at 19.35, 23.235, 37, 85.5 and 91.655 GHz since July 1987" (Page 3, Line 18-20, in the revised manuscript).

Table 3. Summary of the main passive microwave remote sensing sensors.

Sensor	**SSM/I**			**SSMIS**
Satellite	DMSP-F08	DMSP-F11	DMSP-F13	DMSP-F17
On Orbit time	1987-1991	1991-1995	1995-2008	2006-present
Passing Time	A: 06:20 D: 18:20	A: 17:17 D: 05:17	A: 17:58 D: 05:58	A: 17:31 D: 05:31
Frequency & footprint (GHz) : (km × km)		19.35: 45×68 23.235: 40×60 37: 24×36 85.5: 11×16		19.35: 42×70 23.235: 42×70 37: 28×44 91.655: 13×15

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

Response 3: Thank you for your comments. We changed the sentence to the following: "Few user-defined parameters are generally required to optimize the algorithm, such as the number of trees in the ensemble (*n*tree) and the number of random variables at each node (*m*try)" (Page 4, Line 23-24, in the revised manuscript).

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

Response 4: We agree with the reviewer's opinion. In this study, we used the default values of parameters.

5. 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.

Response 5: We agree with the reviewer's opinion. Please refer to the response to "Minor Comment 4" above.

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

Response 6: We agree with the reviewer's opinion. Fig. 7 suggests that the accuracy of the SVM estimation is related to the training data size (Xiao et al., 2018).

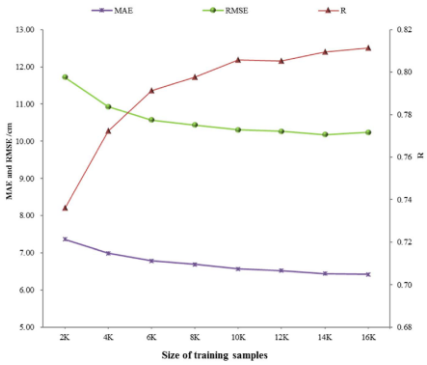

Figure 7. Trend of R (correlation coefficient), MAE (mean absolute error) and RMSE (root mean squared error) with increasing training sample size. K represents one thousand (from Xiao et al., 2018).

In our study, we also analyzed the performances of the RF model with increasing training sample size. The results revealed that the accuracy of RF estimation is insensitive to the training data size (Fig. 8). One of the advantages of the RF model is that it can effectively handle small sample sizes (Biau and Scornet et al., 2016).

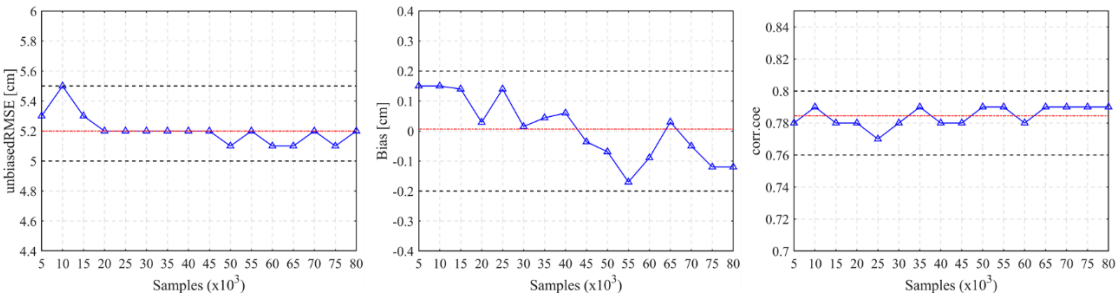

Figure 8. Trends of (a) unbiased RMSE, (b) bias and (c) correlation coefficient with increasing training sample size.

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

Response 7: Thank you for your comments. Please refer to the response to "Minor Comment 6" above.

8. Page 11, lines 4, 5: Which linear model?

Response 8: Thank you for your comments. We have changed the sentence to the following: "The reconstructed product was also compared with the static linear-fitting algorithm developed by fitting 19 and 37 GHz with the snow depth measurements with a constant empirical coefficient over China (Che et al., 2008). The daily snow depth data were obtained from the Environmental and Ecological Science Data Center for West China (http://westdc.westgis.ac.cn) (hereafter, WESTDC product)" (Page 6, Line 17-20, in the revised manuscript).

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

Response 9: Thank you for your comments. We studied whether the machine learning method can overcome the limitations of empirical algorithms. Yang et al. (2019) validated five empirical algorithms and found that this linear model outperformed four other snow depth estimation methods in China. Thus, in this study, we directly compared the estimates of the RF and linear models. We removed this comparison and conducted a more comprehensive analysis of the reconstructed snow depth product.

[1] Yang, J., Jiang, L., Wu, S., Wang, G., Wang, J., and Liu, X.: Development of a Snow Depth Estimation Algorithm over China for the FY-3D/MWRI, Remote Sensing, 11, 977, 10.3390/rs11080977, 2019.

10. Page 12, lines 25–27: This procedure is not clear.

Response 10: We apologize that the description of this procedure was not specific and clear. We omitted this procedure in the revised manuscript according to the reviewers' suggestions. Please refer to the response to "General evaluation 1" above.

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

Response 11: We apologize that the description was not clear. If there are fewer than three available measurements in a pixel during the winter seasons for the 2012-2018 period, the regression coefficients (slope and intercept) can not be calculated. But the snow cover detection method maybe classify this pixel into snow. In such case, we have to assign values to the slope (0.66) and intercept (0) according to the linear model.

We omitted this procedure in the revised manuscript according to the reviewers' suggestions. Please refer to the response to "General evaluation 1" above.

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

Response 12: Thank you for your comment. We added the period in the revised manuscript (Page 4, Line 8-9).

13. 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.

Response 13: We agree with your opinion. Fig. 3 shows that the latitude is highly correlated with the brightness temperature. Thus, latitude has a very slight influence on the predictive performance. However, longitude is poorly correlated with the brightness temperature. Moreover, Fig. 3 indicates that the longitude is more important than latitude to snow depth. We read the reference carefully and cited it as follows: "In addition, RF provides an assessment of the relative importance of predictor variables, which have

proven to be useful for evaluating the relative contribution of input variables (Tyralis et al., 2019b)" (Page 4, Line 29-30, in the revised manuscript).

14. 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.
Response 14: We agree with the reviewer's opinion. Our results also demonstrate that redundant predictor variables slightly affect random forests.

15. Figure 6: Figures should be numbered and respective explanations should be added in the caption.
Response 15: We corrected it.

16. 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.
Response 16: These comments are very useful for improving our paper. We read the reference paper carefully and disscussd the limitaions of the RF model in Section 4.1.
"The RF technique is already used to generate temporal and spatial predictions. Generally, the RF model cannot extrapolate outside the training range (Hengl et al., 2018). Fig. 6 and Table 4 indicate that the spatial predictions of fitted RF algorithms are more biased than are the temporal predictions. Thus, the transferability of a fitted RF algorithm to other areas is in question. Several studies (Prasad, Iverson & Liaw, 2006; Hengl et al., 2017; Vaysse & Lagacherie, 2015; Nussbaum et al., 2018) have proven that RF is a promising technique for spatial prediction; however, these studies aim at spatial prediction of properties that are relatively static over the observational period, e.g., soil types and soil properties.
What makes the Earth system interesting is that it is not static but dynamic (especially concerning snow parameters). Generally, snow depth increases at the beginning of winter and then decreases in spring due to melting. Moreover, snow cover has different spatial patterns in various regions, such as generally deep snow in high-latitude and high-elevation areas. In China, there are five climatological snow classes according to Sturm et al. (1995). Each snow class is defined by an ensemble of snow stratigraphic characteristics, including snow density, grain size, and crystal morphology, which influences the snowpack's microwave signature (Sturm et al., 2010). These dynamic properties of snow will lead to many cases in which the same satellite T_B corresponds to different snow depths, while the same snow depth is associated with various T_B observations, rendering the fitted RF algorithm suboptimal. Using ML techniques in combination with snow forward models (physical modeling) has the potential to overcome many limitations that have hindered a more widespread adoption of ML approaches" (Page 9, Line 22-30, in the revised manuscript).

17. 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.

Response 17: Thank you for your suggestion. We added the information on the RF model (https://cran.r-project.org/web/packages/randomForest): “In this paper, a randomForest R package (Version 4.6-14) is used for regression (Liaw and Wiener 2002; Breiman et al. 2018)” (Page 5, Line 1-2, in the revised manuscript).

Language

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

Response 1: We agree with your opinion. We changed “prediction” to “regression” in the revised manuscript.