Response to Reviewer Comments by Divyesh Varade 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 all been very helpful for revising and improving our paper as well as providing important guidance for our research. We have studied the comments carefully and have made corrections, which we hope meet with approval. The detailed corrections and the responses to your comments are listed below point by point:

Review #1

General Comments: Snow depth estimates are significant for the assessment of the hydrological potential of the snowpack. The application of machine learning tools provides us with a means to derive new depth estimates from a trained model. The methods for the modeling of snow depth using remote sensing data are predominantly based on passive microwave data with much higher repeatability and spatial coverage than InSAR data, rendering such analysis suitable for the monitoring of the snow accumulation. I thus, consider this work to be significant.
Overall, the manuscript is organized and written neatly and represented in a well-structured manner. The language is mostly appropriate except for a few sentences which are not easily understandable. There are some claims and statements made by the authors that lack references or evidence. This work is appreciable in the extent of the analysis performed by the authors, in particular for the time series evolution of the snow depth in some of the major provinces in China. However, the manuscript also presents some weaknesses in the methodology, experiments, and particularly the validation.

Specific comments:

1. The authors have not clearly stated the novelty of their proposed method. In my opinion, the novelty of the proposed method is in the design of the regression model using the Random Forests i.e. the step -1 in Figure 3 and its application for the modeling of snow depth. The other steps are similar to the methodology proposed in – Jiang, L., Wang, P., Zhang, L. et al. Sci. China Earth Sci. (2014) 57: 1278. https://doi.org/10.1007/s11430-013-4798-8.

Response 1: Thank you for your comments, we agree on your original assessment of novelty, and this point was indeed weakly represented in the original manuscript. However, we have now redesigned the methodology in order to further increase the novelty with respect to previous studies. Specifically, there are now four RF algorithms trained with different predictive variables. Temporally and spatially independent datasets were used to validate the fitted RF algorithms. The aims were to

(1) test whether certain choices of predictive variables are necessary and whether they improve the RF algorithm;
(2) demonstrate the transferability in spatial and temporal scales.

We rewrote the part of the introduction concerning novelty, and it now reads as follows: "The primary objectives of this study are to assess the feasibility of the RF model in estimating snow depth, to determine whether the inclusion of auxiliary information (geolocation, elevation and land cover fraction) contributes to the improvement of RF, and eventually to develop a time series (1987 to 2018) of snow depth data in China and analyze the trends in annual mean snow depth. To complete the feasibility study of the RF model, we designed four RF algorithms trained with different combinations of predictor variables and validated them using temporally and spatially independent reference data. To our knowledge, this type of assessment of RF algorithm performance has not been made to date over China" (Page 3, Line 7-11, in the revised manuscript).

2. Why the Random Forest is used, in contrast to better alternatives such as deep neural networks? The authors claim that RF is superior to SVM and ANN, is there any documented evidence regarding RF to be superior to SVM or ANN in link with modelling of geophysical parameters similar to snow depth? Deep learning for classification and regression has been found very useful in recent literature. What is the reason that the authors use RF instead of deep neural networks? Please provide evidence for this or perform additional experiments to prove that RF-based estimates are superior to SVM, ANN, and deep NN based estimates.

Response 2: Thank you for your comments. In our view, any machine learning model has both advantages and disadvantages. Over the last two decades, RF has been one of the most successful machine learning algorithms for practical applications due to its proven accuracy, stability, speed of processing and ease of use (Reichstein et al., 2019). Thus, we studied whether the RF model could be used to retrieve snow depth in this study.

We also conducted a comparison between RF and ANN. The training data were from the training stations during the period 2012-2014 (Fig. 2). The predictor variables included brightness temperatures (19 GHz and 37 GHz at vertical polarization), latitude, longitude, elevation and land cover fraction. We used spatially independent data from validation stations (2015-2018) to verify the fitted ANN and RF algorithms. The results showed that the RF model was superior to ANN with respect to snow depth estimation in China (Fig. 1).

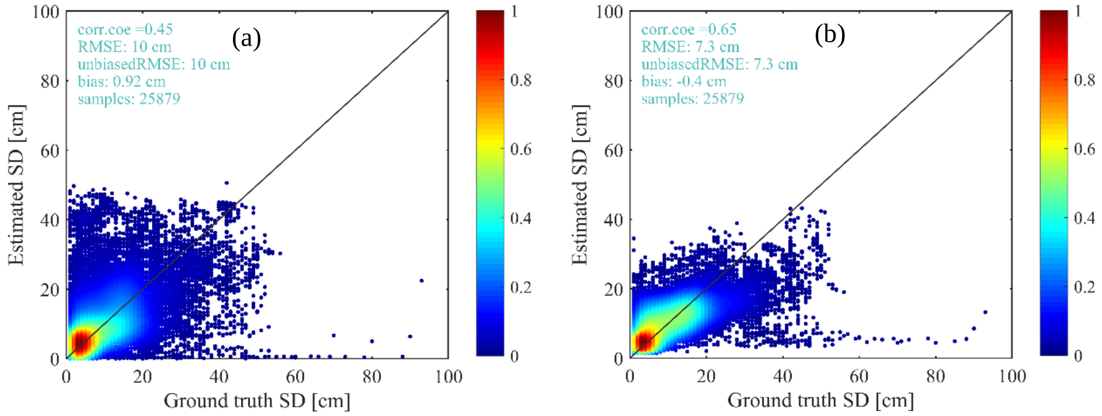

Figure 1. Comparison between (a) ANN and (b) RF with respect to snow depth estimation in China.

As you pointed out, there are a few pitfalls such as the risk of naive extrapolation and poor transferability in spatially limiting the applications in spatio-temporal dynamics. It is in this realm that the techniques of deep learning promise breakthroughs. We are attempting to operate the Deep Neural Networks (DNN) model to overcome the limitations of traditional machine learning approaches.

[1] Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat.: Deep learning and process understanding for data-driven Earth system science, Nature 566, 195–204, 2019.

We also rewrote the sentence, and now it reads as follows: "Over the last two decades, RF has been one of the most successful ML algorithms for practical applications due to its proven accuracy, stability, speed of processing and ease of use (Rodriguez-Galiano et al., 2012; Belgiu et al., 2016; Maxwell et al., 2018; Bair et al., 2018; Qu et al., 2019; Reichstein et al., 2019, Tyralis et sl., 2019a)" (Page 3, Line 2-5, in the revised manuscript).

3. In both cases, steps 1 and 3, the authors use only a single year data for validation. This neither provides enough points for validation nor any comprehensive inferences from the validation results.

Response 3: We are sorry for the confusion. The term (2012-2013) refers to two years of data, not single year. However, it does not matter because we have redesigned the methodology and added more validation data. Available stations were randomly divided into two roughly equal-sized parts by Matlab software (Fig. 2). The data from training stations (Fig. 2) during the period 2012-2014 were used to train the RF model. The dataset from validation stations during the period 2015-2018 was used to assess the accuracy of the fitted RF algorithm.

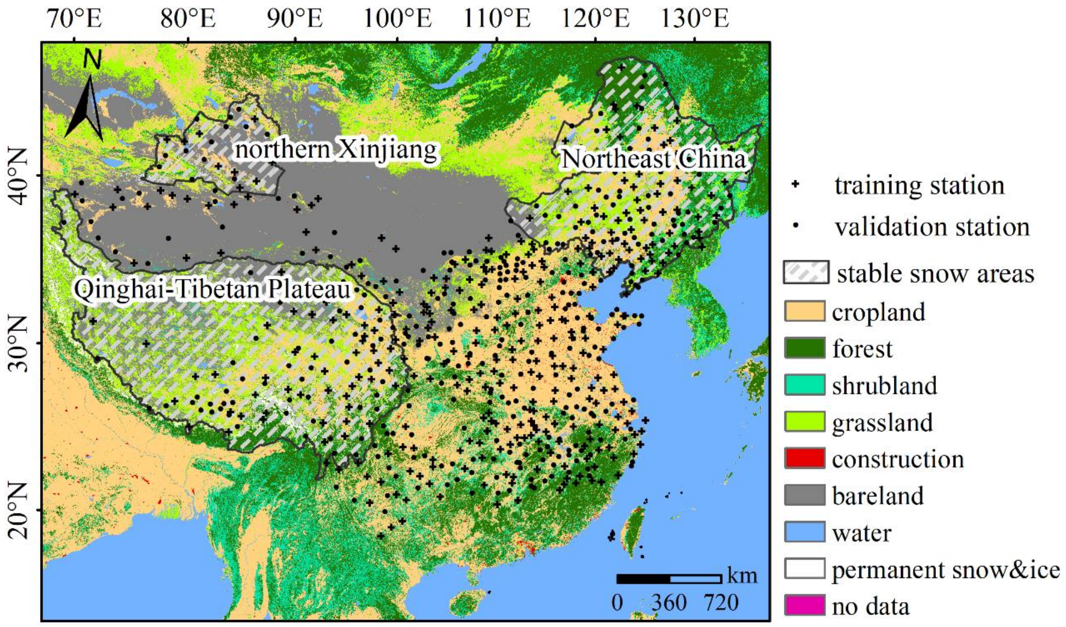

Figure 2. Spatial distribution of the weather stations and land cover types in the study area. There are three stable snow cover areas in China: Northeast China (NE), northern Xinjiang (XJ) and the Qinghai-Tibetan Plateau (QTP).

In this study, we used the fitted algorithm to reconstruct a long-term snow depth dataset (1987 to 2018) directly. Then, this product was evaluated by the independent ground truth measurements over the period 1987-2018 from the validation stations (Fig. 3) and was also compared with the former snow depth data (WESTDC) in China (Fig. 4).

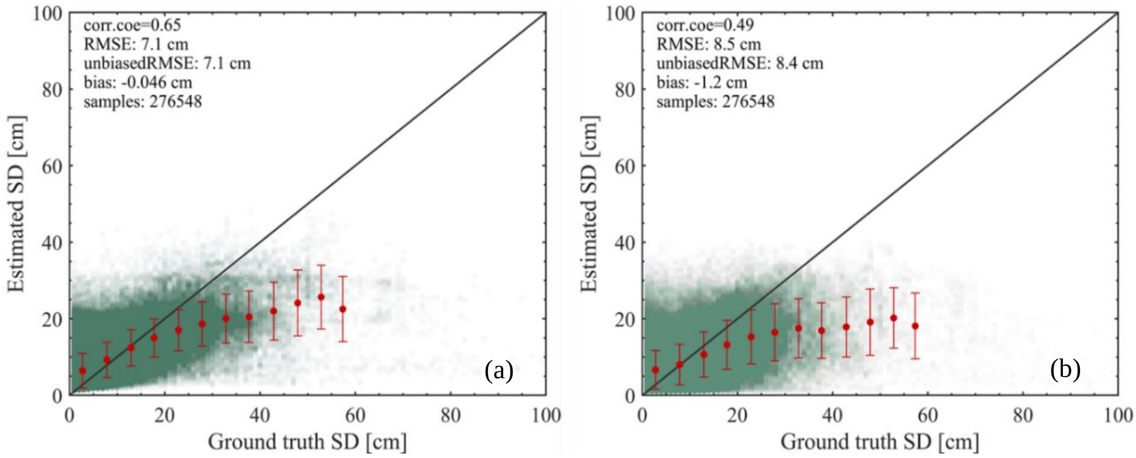

Figure 3. Scatterplots of the estimated snow depth and the ground truth observation for (a) RF and (b) WESTDC products.

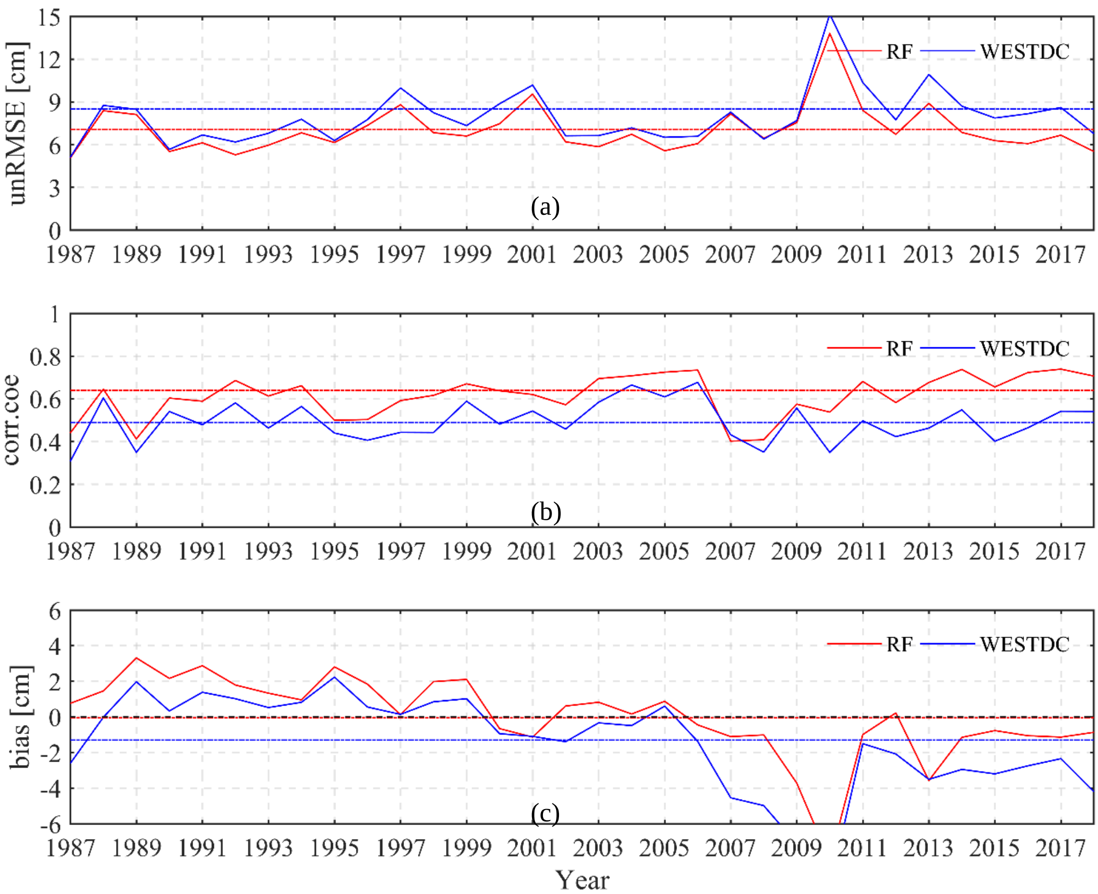

Figure 4. Time series of (a) unbiased RMSE (unRMSE), (b) correlation coefficient (corr.coe) and (c) bias for RF and WESTDC products. The colorful dashed lines represent mean values of assessment indexes.

4. The datasets used for training and testing have some issues. The authors have shown how the actual depth has varied through the years 1987-2019. But for training only data till 2004 was used. The trends from Figures 10 and 11 show a marginal decrease in the mean snow depth. Would it not be better to use data from every two year or alternate year for training the RF. Similarly for testing, the authors use data from the only year 2012-13 for model testing and 2017-18 for testing the final results. This is not sufficient to develop a comprehensive interpretation of the results.

Response 4: Thank you for your comments. We indeed have collected ground truth snow depth observations from 1987-2018. To determine the appropriate number of training samples, a test was conducted to analyze the sensitivity of the RF model to training sample size. To ensure there were enough samples, we selected 80,000 samples from 1987 to 2004 as available training data, and a two-year dataset from 2005 to 2006 was applied to assess the performance.

We agree with your opinion regarding the validation using much more data, and these comments are very constructive. Thus, we have added more data to validate the fitted RF algorithms and the reconstructed snow depth product. Please refer to the response to "Specific comment 4" above.

5. In section 3.2, the correlation coefficient is 0.77. Is this satisfactory enough to be used to generate the reference dataset from the RF model? A majority of data are below 10 cm snow depth, then an error of 4.5 cm is significantly high. To have a better understanding of the modeled results, it is vital that we observe the accuracy for the points of higher snow depth also. Particularly, when there is a very high snow depth different for the regions QTP and the others. The validation should be carried out for these regions separately. I suggest the authors show a histogram of the data and also carry out a separate fit for points of snow depth >10cm or perform a case by case fit with respect to the study area. A significant concern is that in the case of shallow snow (<10cm), is the brightness temperature actually representative of the contributions from the shallow snowpack or the underlying ground. This requires further investigations. This is important since the bulk of the data is within the 0-10 cm range. Another concern is that there are very few points with snow depth >40cm. In several locations in the Himalayas, the peak snow depth is usually around 1m or more. Thus, the applicability of the proposed method or the transferability of the proposed method to other areas, in these cases, is in question.

Response 5: Thank you for your comments. Other reviewers gave similar comments. Since the dataset obtained by RF is an approximation of the true snow depth, the uncertainties are transferred to the second stage of prediction. Other reviewers suggested that we directly use the fitted RF algorithm to produce the long-term snow depth data in the period 1987-2018.

Figure 5 shows the histograms of 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. 5.

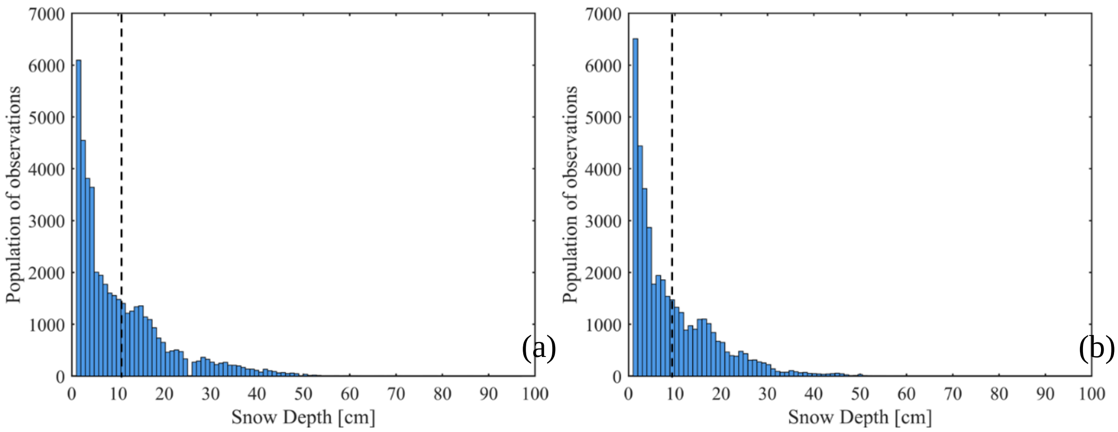

Figure 5. Histograms of snow depth observations from (a) training and (b) validation stations. The average values (black dashed lines) are equal to 10.5 cm and 9.8 cm, respectively.

The idea to carry out a separate fit for points of snow depth > 10 cm is good, but it cannot be used to estimate snow depth in space and time. This is because passive microwave observations cannot distinguish deep and shallow snow cover so that the background of snow depth is unknown. Thus, for a snow cover satellite pixel, we don't know which fitted RF algorithm should be used to retrieve snow depth.

We agree with your comments about underestimations for deep snow. The validation was carried out for three snow cover regions in China separately (Fig. 6).

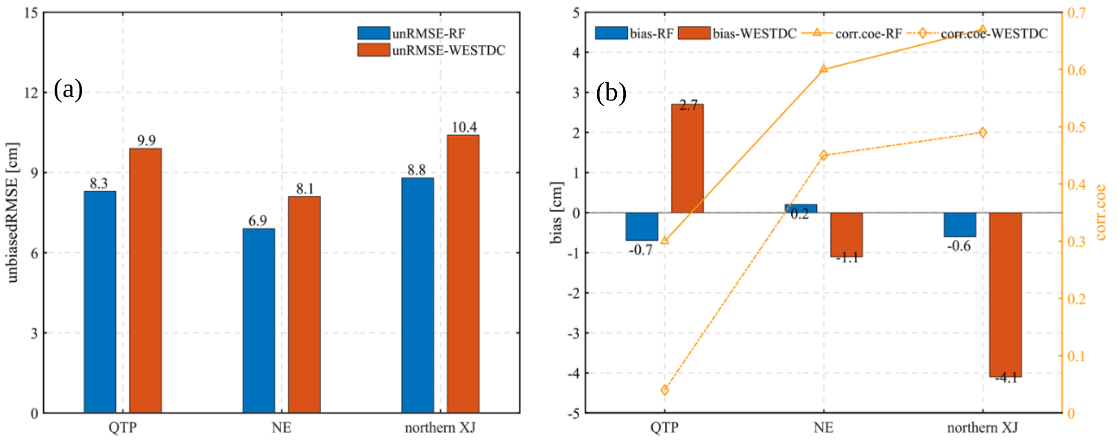

Figure 6. The validation of RF and WESTDC snow depth products in three stable snow cover areas in China with respect to (a) the unbiased RMSE, (b) bias and correlation coefficient.

We selected 20 cm as a threshold to assess the performances in deep (> 20 cm) and shallow (≤ 20 cm) snow cover. The percentage of shallow snow conditions to total samples was approximately 90%. Table 1 displays the comparison between RF estimates and WESTDC product in the three snow cover areas. Both products presented a notable underestimation for deep snow cover, with the biases of -34.1 cm and -33.8 cm in QTP for the RF and WESTDC products, respectively. The biases were -10.4 cm and -8.9 cm for the RF product in NE and northern XJ, respectively, whereas they were -11.8 cm and -13.2 cm for the WESTDC data. For shallow snow cover, the RF product is superior to the WESTDC estimates in QTP, with unbiased RMSEs of 3.4 cm (RF) and 5.6 cm (WESTDC).

Furthermore, the WESTDC product presents an overestimation in QTP, with a bias of 4.0 cm that is much higher than the RF's 0.6 cm.

Table 1. Comparison between RF estimates and WESTDC product in three stable snow cover areas for deep (> 20 cm) and shallow (≤ 20 cm) snow cover.

RF product						
Regions	QTP		NE		northern XJ	
SnowDepth (cm)	<= 20	> 20	<= 20	> 20	<= 20	> 20
corr.coe	0.30	0.06	0.49	0.17	0.48	0.31
bias (cm)	0.59	-34.12	1.79	-10.38	2.52	-8.85
unRMSE (cm)	3.43	20.70	5.36	7.00	6.12	9.62
Samples	15503 (96.4%)	583 (3.6%)	151939 (87.3%)	22168 (12.7%)	32468 (69.8%)	14051 (30.2%)
WESTDC product						
Regions	QTP		NE		northern XJ	
SnowDepth (cm)	<= 20	> 20	<= 20	> 20	<= 20	> 20
corr.coe	0.16	-0.18	0.37	0.03	0.34	0.16
bias (cm)	4.02	-33.78	0.47	-11.75	-0.39	-13.22
unRMSE (cm)	5.60	21.62	6.47	9.10	7.35	11.30
Samples	15503 (96.4%)	583 (3.6%)	151939 (87.3%)	22168 (12.7%)	32468 (69.8%)	14051 (30.2%)

We presented the potential errors of the reconstrued snow depth in Section 4.3 as follows: "Fig. 7 indicates that the RF model does not fully solve the overestimation and underestimation problems. For deep snow (> 20 cm), the biases are up to -8.9 cm and -10.4 cm in NE and northern XJ, respectively. Deep snow conditions account for roughly 10% of all training samples (Fig. 2). The estimates for deep snow cover in the QTP exhibit a large bias of -34.1 mm. Fig. 6 also illustrates that the fitted RF algorithms have no predictive ability for extremely deep snow conditions, especially in QTP. We checked the training data and found that the extreme high snow depth data (> 60 cm) occurred in QTP. However, the number of such cases is very small. In addition, the station measurements are point values while the satellite grids have a spatial resolution of 25 km × 25 km. Thus, the representativeness of these data is questionable. Snow depth estimation in the mountains remains a challenge (Lettenmaier et al., 2015; Dozier et al., 2016; Dahri et al., 2018). Numerous studies have been conducted on the snow cover over the QTP and have indicated that the snow cover in the Himalayas is higher than elsewhere, ranging from 80% to 100% during the winter (Basang et al., 2017; Hao et al., 2018). Additionally, Dai et al. (2018) showed that deep snow (greater than 20 cm) was mainly distributed in the Himalayas, Pamir, and Southeastern Mountains. Thus, the RF product produced in this paper has poor performance in QTP for the deep snow cover.

Table 5 indicates that there is overestimation in NE and northern XJ for shallow snow cover, which may be due to the following reasons. First, the PMW signals are insensitive to thin snow cover, especially for fresh snow with low snow density and snow grain size. Second, the large diurnal temperature range tends to subject the snowpack to frequent freeze-thaw cycles and leads to rapid snow grain (~2 mm) and snow density (200-350 kg/m^3) growth and consequently a high T_B difference (Meløysund et al., 2007; Durand et al., 2008; Yang et al., 2015; Dai et al., 2017). Third, frozen soil reduces the accuracy of estimates. Both snow and frozen ground are volume-scattering materials, and they have similar microwave radiation characteristics, making them difficult to distinguish. In addition, a limiting factor in estimating snow depth for PMW remote sensing is the presence of liquid water. In this study, a snow cover detection method is used to filter out wet snow cover; however, there are still misclassification errors, especially at the end of the winter season (Grody and Basist., 1996; Liu et al., 2018). In such cases, satellite observations are mainly associated with the emissions from the wet surface of the snowpack. Therefore, in wet snow conditions, snow depth retrieval is not possible (Derksen et al., 2010; Tedesco et al., 2016)" (Page 10, Line 19-28, Page 11, Line 1-13, in the revised manuscript).

6. The authors observed higher errors for shallow snow depth, but the manuscript lacks any discussion on the contributions from the underlying ground layer to the passive microwave brightness temperature in case of shallow snow depth. The authors have simply added some references. A discussion is required in the manuscript on the sensitivity of snowpack thickness and stratigraphy towards the passive microwave brightness temperature.

Response 6: We redesigned the methodology in this study. The new RF product presented lower errors under shallow snow cover conditions (Table 1). We have discussed this finding in Section 4.3. Please refer to the response to "Specific comment 5" above.

The microwave emission model of layered snowpack (MEMLS) was applied to simulate the T_B with varying snow parameters (Mätzler et al., 1999; Löwe et al., 2015; Pan et al., 2015). Fig. 7 shows the sensitivity of snow depth to T_B at 36 GHz for various snow density and snow grain size. Generally, the snow density (< 100 kg/m^3) and snow grain size (correlation length < 0.2 mm) are small for shallow snow cover (< 5 cm). The passive microwave signals are insensitive to the shallow snow cover. Moreover, the snow cover is patchy under shallow snow conditions, challenging the relationship between satellite T_B and snow depth.

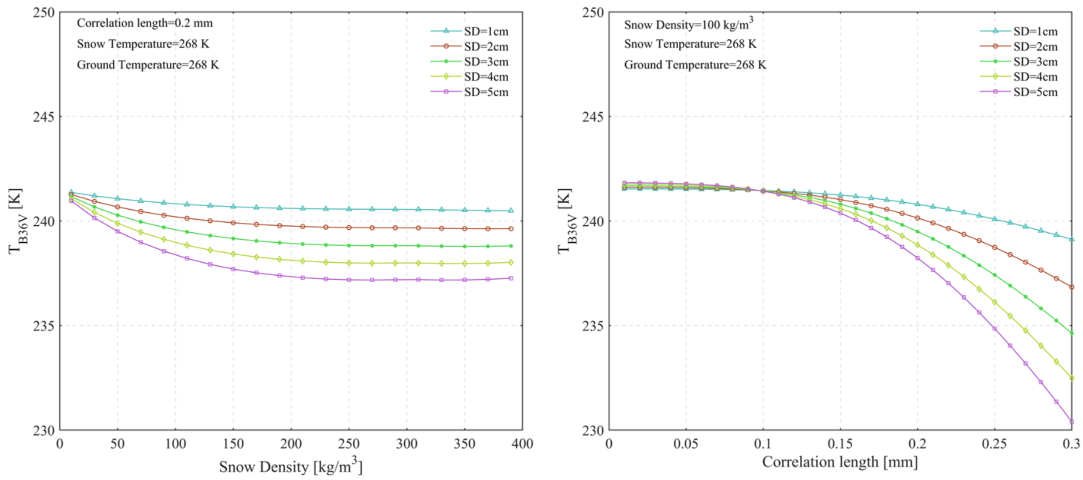

Figure 7. The sensitivity of snowpack stratigraphy to the passive microwave brightness temperature simulated with the MEMLS model.

7. Page 12, L25-27: Does this mean 3-10 samples in (3x25)x(3x25) sq. km area? This is not clear to me. I think the authors are referring to measurements from field campaigns or weather stations as samples. In this case, the number of samples is very small per the averaging window. Please provide references for this.

Response 7: Thank you for your comments. We apologize that the description of this part was not clear. We have redesigned the paper and removed the pixel-based method according to other reviewers' comments.

8. Figures 9a and 9b. There are very few samples used for validation in these figures. Further, these samples are discontinuous (Figure 9a) and therefore, this should not be used as the basis for ascertaining the performance of the proposed method, since due to the distribution of the points, it is expected that the fit will provide better results.

The authors may perform other significance tests such as Nemar's test, but the fact remains that the validation data is not really comprehensive. The data shown in Figure 9b is much better for assessment, as it is continuous. But why only 10 points? Earlier it was shown that several ground stations exist in the area. I suggest the authors also use data from other years in their validation scheme, as the results shown at present are not convincing. Why is the modeled snow depth showing very less sensitivity between 20-40cm (nearly constant) and again afterward? This is an issue that requires investigation.

Response 8: Thank you for your constructive comments. We used independent ground truth observations from 1987 to 2018 to validate the RF product. Fig. 3 shows the error bars and scatterplots. The "o" marker is the mean snow depth computed at each corresponding ground truth bin, while upper and lower colorful bars indicate one standard deviation from the mean. There are almost 280,000 samples. Please refer to the response to "Specific comment 3" above.

9. In section 4.5, the selection of sample size for training and testing is reversed. Since the MEMLS requires auxiliary information, which is seldom available, the training samples should be much less than the validation samples. This validation strategy is not convincing. From the discrepancy in the training and testing samples, it is already expected that the model accuracy would be high.

Response 9: We appreciate your suggestions. The aim of this part work is to demonstrate that more prior snow information can improve the performance of the RF model. Reviewer #4 suggested we should omit this part and return to the combination in a future publication. Thus, combining the snow forward model with the ML method will be the focus of our future work.

Minor issues:

1. Page 02, L7: " the Himalayas during: : :.". The Himalayan ranges are very long and are shared by several countries. Please specify which Himalayan ranges the authors are referring to here. I do not agree with the statement that mean snow depth is maximum in Xinjiang for the entire Himalayan range. Please provide references for this.

Response 1: We apologize for the confusion. Three snow cover areas are shown in Fig.1 (Please refer to the response to "Specific comment 3" above). The trend analysis of snow depth was conducted based on the ground truth observations, RF dataset and WESTDC product during the period 1987-2018. To illustrate the different changing patterns, the trends in northern XJ, NE and QTP were analyzed.

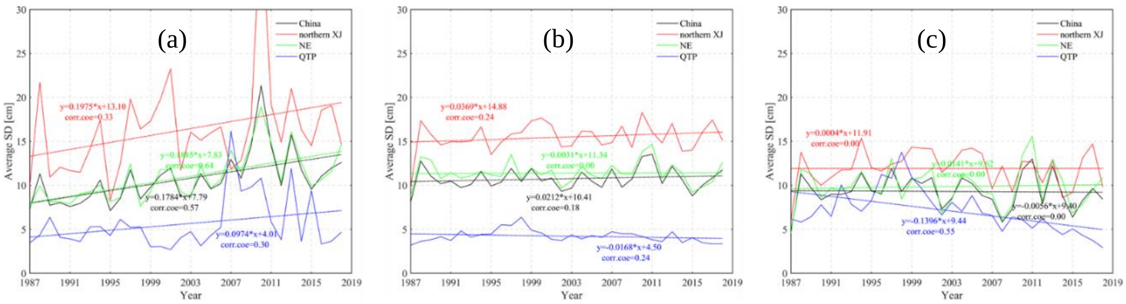

Figure 8. The trend analysis of snow depth based on (a) station observations, (b) RF estimates, and (d) WESTDC product in three stable snow cover areas in China. The correlation is statistically significant at the 0.05 level.

We rewrote the sentence as follows: "On a temporal scale, the ground truth snow depth presented a significant increasing trend from 1987 to 2018, especially in NE. However, the RF and WESTDC products displayed no significant changing trends except in QTP. The WESTDC product presented a significant decreasing trend in QTP, with a correlation coefficient of -0.55, whereas there were no significant trends for ground truth observations and the RF product" (Page 1, Line 26-29 in the revised manuscript).

2. Page 02: L8-11: These are documented facts in literature for several other locations, however. Thus, the authors should strictly restrict their inferences to their own findings and not speculate. Thus, here the sentence should be specific to the study area in the manuscript.

Response 2: We appreciate your suggestions. Three snow cover areas in China are shown Fig. 1. The time series of mean snow depth in three stable snow cover areas over China is shown in Fig. 8. Fig. 8a shows that the mean snow depth in northern XJ is the largest among the three regions, and the pattern in NE is highly consistent with the overall trend in China. Comparing the ground truth data and RF product (Fig. 8a *vs*. 8b) shows that there are similar patterns in terms of the magnitude of snow depth in the three snow cover areas.

3. Page 02, L11-13: The sentence "In conclusion: : :." is not clear. Please rephrase.

Response 3: We consider this sentence to be unnecessary and have removed it.

4. Page 02, L24: "mean snow density". I believe the authors are here referring to mean stratigraphic snow density". Please correct this.

Response 4: Thank you for your comments. Reviewer #4 thought this paper should focus on snow depth and not snow water equivalent. Thus, we removed this description and rewrote the sentence as follows: "Snow depth is a crucial parameter for climate studies, hydrological applications and weather forecasts (Foster et al., 2011; Takala et al., 2017; Tedesco et al., 2016; Safavi et al., 2017)" (Page 2, Line 4-6, in the revised manuscript).

5. Page 03, L17-18: "however, these: : :". Is there any evidence that the RTM based methods are computationally more expensive than machine learning-based methods. In my opinion, both depend on the selection of the parameters. For example, an RF with substantial input and a high number of trees may be as expensive computationally. If there is no documented evidence on this, please remove this statement.

Response 5: We deleted the sentence in the revised manuscript.

6. Page 11, L 11-13: Please correct the range as 200-350 kg/m3 and provide a reference, for example- Meløysund, Vivian, Bernt Leira, Karl V. Høiseth, and Kim R. Lisø. 2007. “Predicting snow density using meteorological data.” Meteorological Applications14 (4): 413–23. doi:10.1002/met.40.

Response 6: We appreciate the reviewer’s help and suggestions. We read the reference carefully. It is a good paper and very useful for us. We corrected the range and cited the reference in the revised manuscript (Page 11, Page 5-7).

7. Page 17, L20: “The snowpack is set ..”. This should be the snowpack is assumed to comprise a single layer indicating a semi-infinite medium. This is a common assumption in electromagnetic modeling of the snowpack. Please add references to this.

Response 7: We removed this part. Please refer to the response to “Specific comment 9” above.

8. Figure 1: This needs to be revised. Firstly, the authors use 3 areas for their study which have not been shown on the large map. Secondly, the two pixels mentioned previously should be shown at a higher resolution. Third, write in captions what the color bar represents, is it elevation? Finally, the pixels shown should also have a lat-long grid and scale bar.

Response 8: We appreciate the reviewer’s comments and suggestions. We redesigned the map (Fig. 1). Because of the paucity of samples from the field sampling campaign, we omitted these data and added more station observations (1987 to 2018) as a new validation dataset.

9. Figure 7: Why is the number of points and their locations changing in the maps showing stations. I believe this should remain fixed irrespective of the month. If there is no snow at some of the stations which have been omitted, these should be shown with either a different symbol or a color.

Response 9: As you pointed out, the number of available station observations is not fixed during the snow winter season. In the revised manuscript, we have deleted this statement.

10. Figure 8: The images are distorted. It appears as if they were stretched manually to fit some size.

Response 10: Thank you for your comments. The pixel-based algorithm was omitted in the revised manuscript. Please refer to the response to “Specific comment 5” above.

11. Figure 9/Table 4 and several other instances: The R2 and R, i.e. the determination coefficient and the correlation coefficient, respectively, are two different parameters and have been used interchangeably with similar symbols in the manuscript, which makes it difficult to judge the accuracy of the results.

Response 11: We apologize that we did not describe this consistently. We corrected it in the revised manuscript.

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.

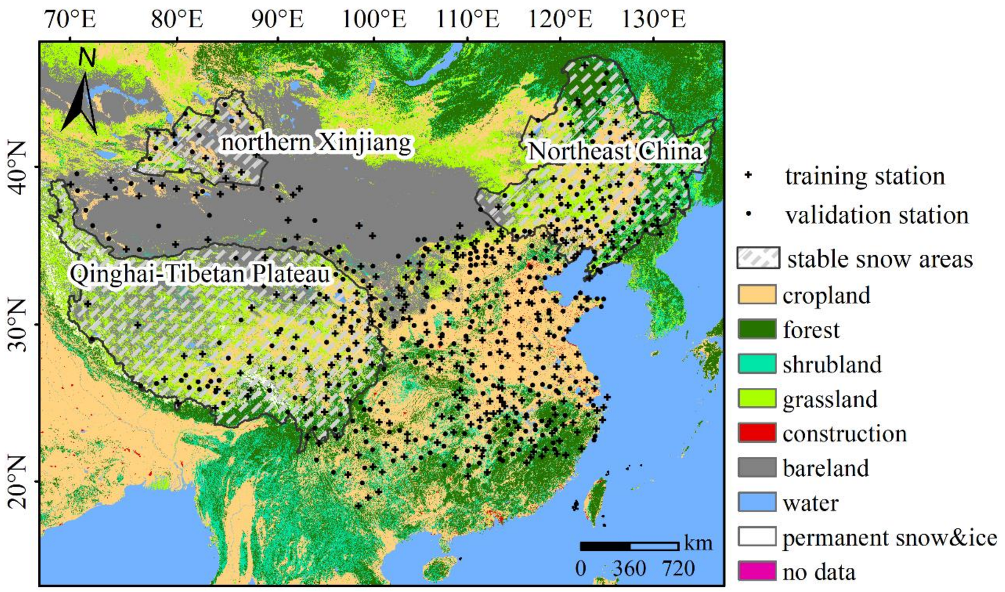

Figure 1. Spatial distribution of the weather stations and land cover types in the study area. There are three stable snow cover areas in China: Northeast China (NE), northern Xinjiang (XJ) and the Qinghai-Tibetan Plateau (QTP).

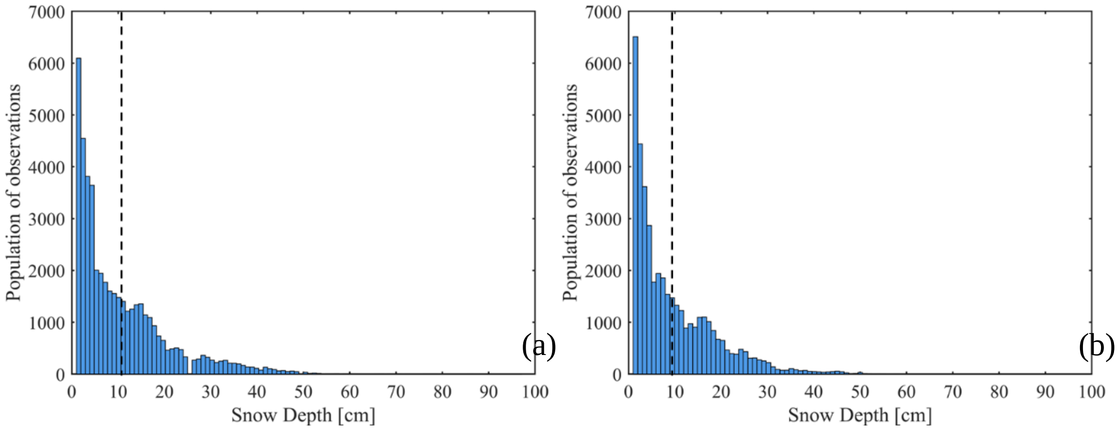

Figure 2. Histograms of snow depth observations from (a) training and (b) validation stations. The average values (black dashed lines) are equal to 10.5 cm and 9.8 cm, respectively.

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.

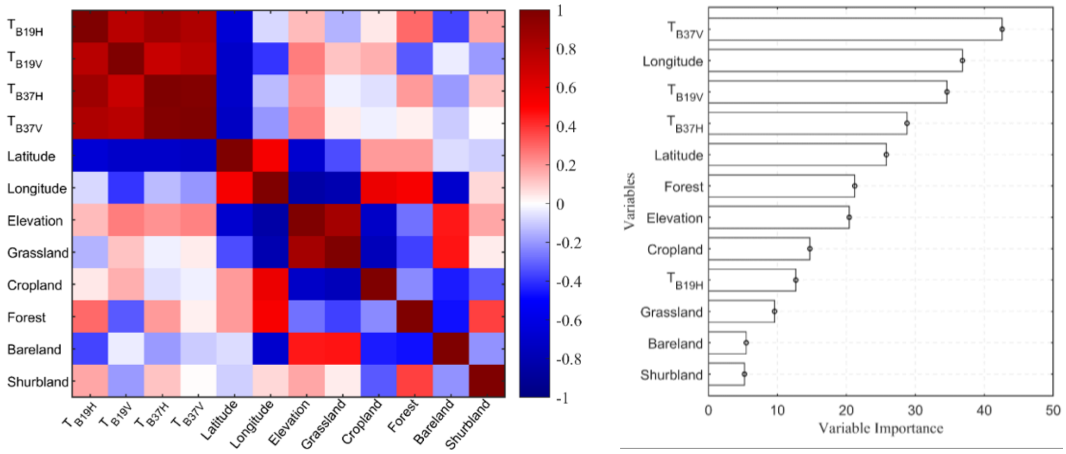

Figure 3. Correlations between the predictor variables (left) and the ranking of variable importance (right). The importance of variables, referred to as Mean Decrease Accuracy (MDA) in the RF model, is obtained by averaging the difference in out-of-bag error estimation before and after the permutation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

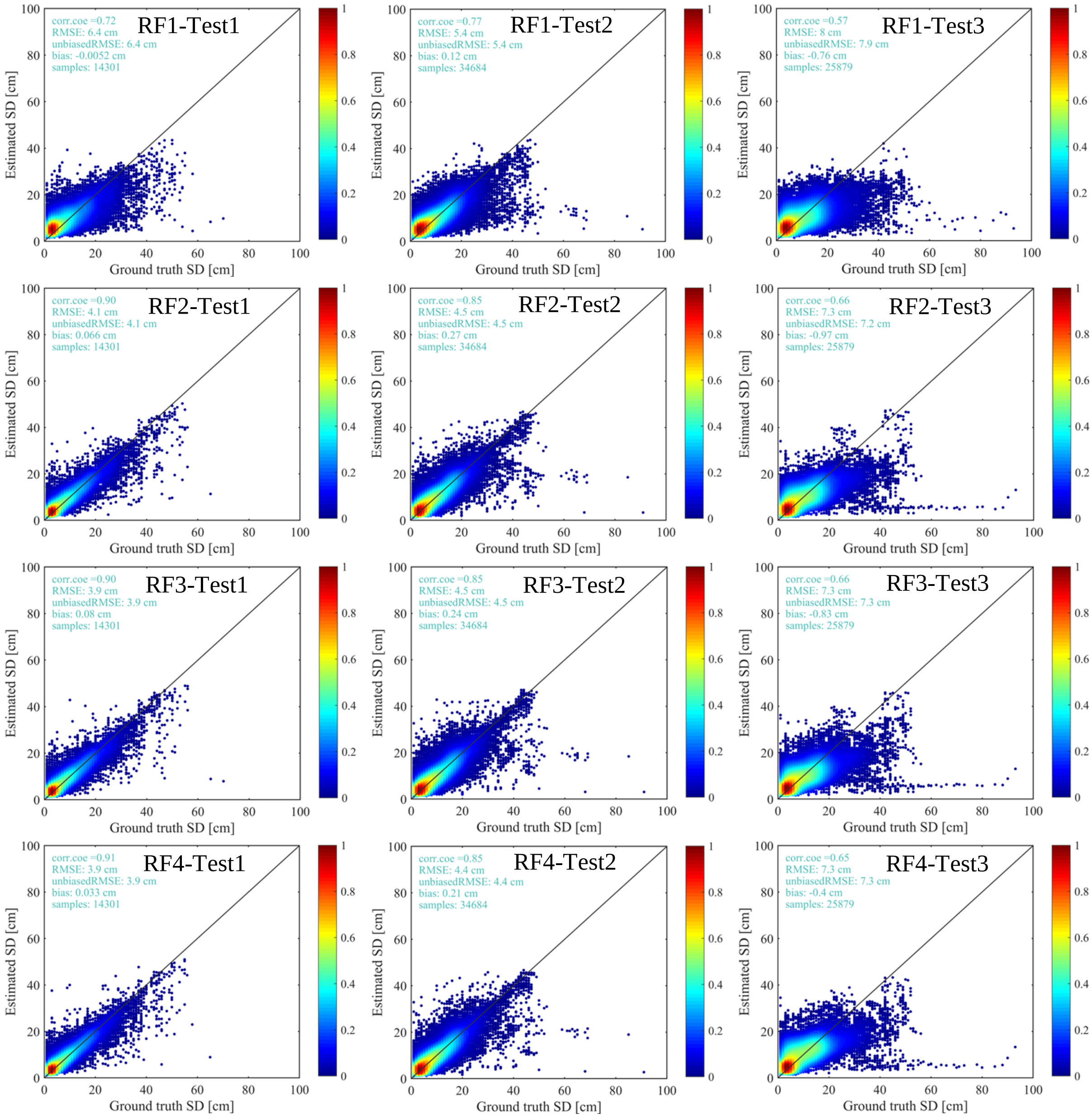

Figure 4. The color-density scatterplots of the estimated snow depth with four fitted RF algorithms and the ground truth snow depth. The four trained RF algorithms (RF1, RF2, RF3, RF4) were evaluated with three validation datasets (Test1, Test2, Test3).

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).

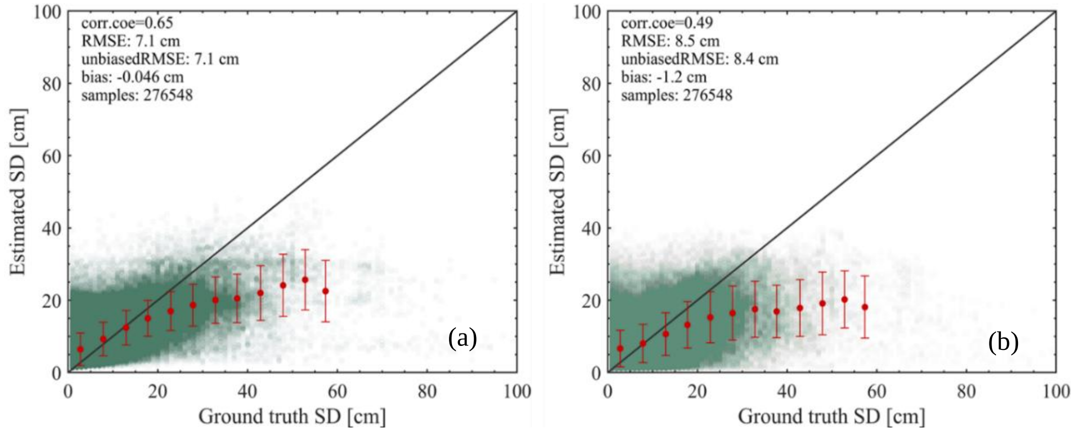

Figure 5. Scatterplots of the estimated snow depth and the ground truth observation for (a) RF and (b) WESTDC products.

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.

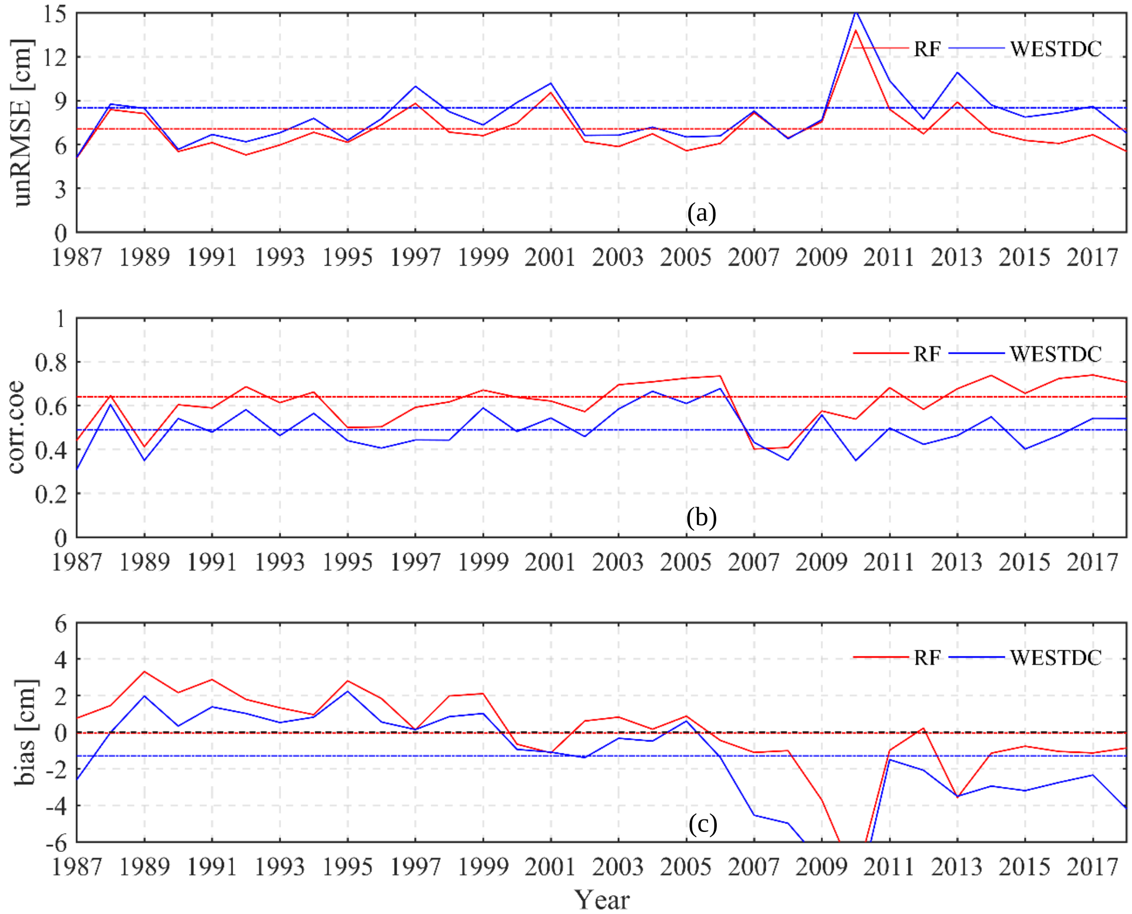

Figure 6. Time series of (a) unbiased RMSE (unRMSE), (b) correlation coefficient (corr.coe) and (c) bias for RF and WESTDC products. The colorful dashed lines represent mean values of assessment indexes.

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).

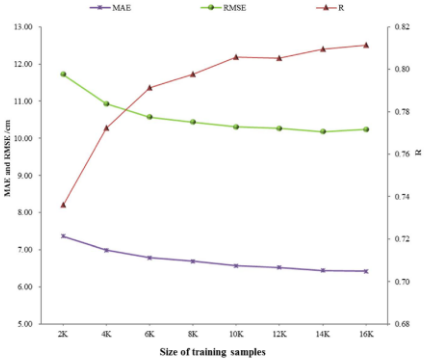

Figure 7. Trend of R (correlation coefficient), MAE (mean absolute error) and RMSE (root mean squared error) with increasing training sample size. K represents one thousand (from Xiao et al., 2018).

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).

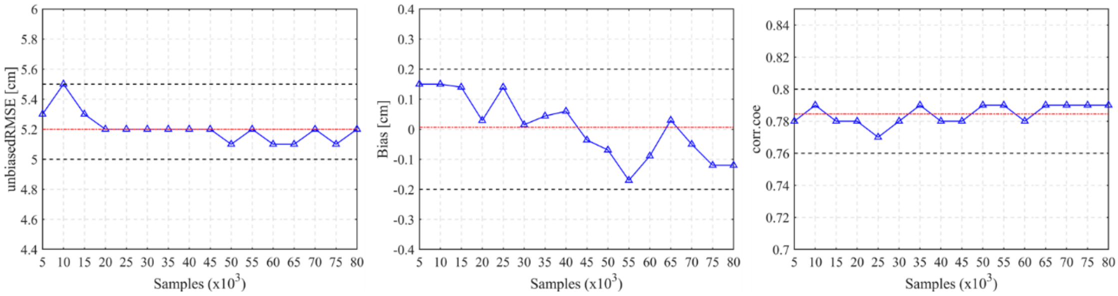

Figure 8. Trends of (a) unbiased RMSE, (b) bias and (c) correlation coefficient with increasing training sample size.

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.

Response to Reviewer Comments by Tomasz Berezowski 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 #3

General Comments: The manuscripts aims to reconstruct the historical snow data set and to develop a real time snow depth estimation. I qualify this manuscript somewhere between major revision and rejection. The major revision is because the MS has some serious issues in methods, validation and some of the statements are not supported by the result. On the other hand the historical snow data set is an interesting product (if properly validated). The rejection is due to lack of novelty in this study: Authors use well established methods in a standard way and what they obtain is a product that has a similar RMSE as a former product available for China.

Response: Thank you for your comments. We revised the manuscript carefully and thoroughly. According to yours and other reviewers' suggestions, we redesigned the methodology and conducted the comparisons between the complicated and simple methods to demonstrate which procedure is more effective for snow depth estimation, also improving novelty of the study. The primary objectives of this study are to assess the feasibility of the RF model in estimating snow depth, to determine whether the inclusion of auxiliary information (geolocation, elevation and land cover fraction) contributes to the improvement of RF, and eventually to develop a time series (1987 to 2018) of snow depth data in China and analyze the trends in annual mean snow depth. To complete the feasibility study of the RF model, we designed four RF algorithms trained with different combinations of predictor variables and validated them using temporally and spatially independent reference data. To our knowledge, this type of assessment of RF algorithm performance has not been made to date over China. The reconstructed snow depth dataset is now available and we will upload it later. There are four major revisions in this study.

1) **Revision 1: scientific validation dataset**

One of major issues in the original manuscript was the validation data are not temporally and spatially independent. Thus, in the revised manuscript, available stations in China 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.

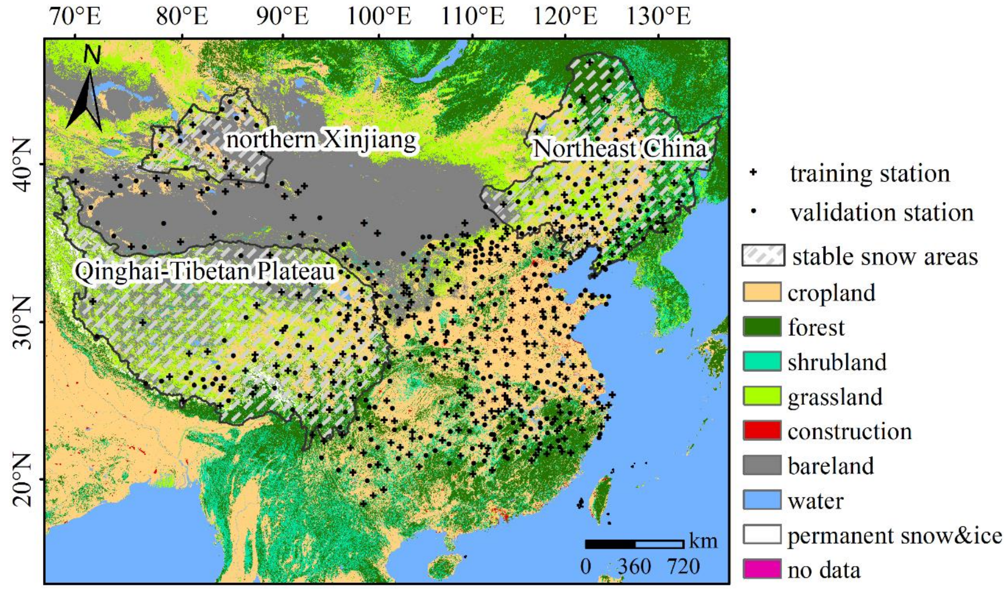

Figure 1. Spatial distribution of the weather stations and land cover types in the study area. There are three stable snow cover areas in China: Northeast China (NE), northern Xinjiang (XJ) and the Qinghai-Tibetan Plateau (QTP).

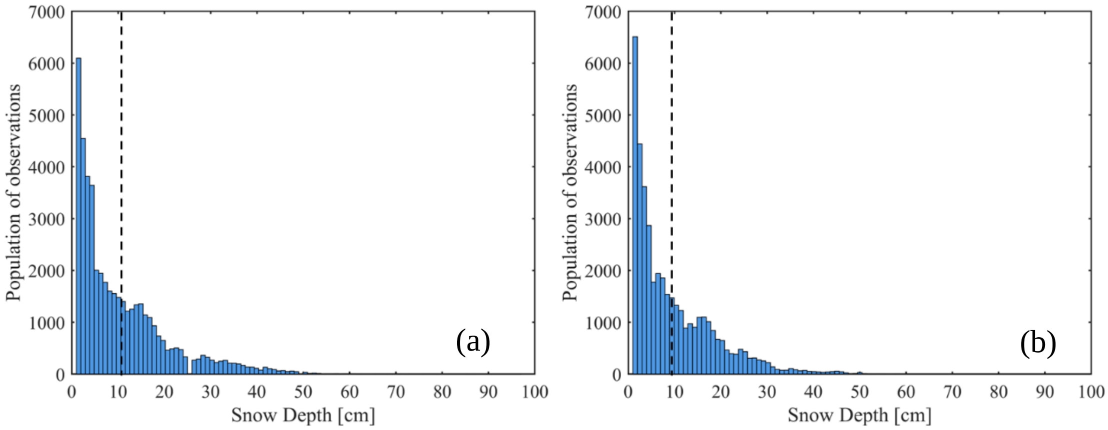

Figure 2. Histograms of snow depth observations from (a) training and (b) validation stations. The average values (black dashed lines) are equal to 10.5 cm and 9.8 cm, respectively.

2) Revision 2: four selection rules of predictor variables

The procedure described in the original manuscript was 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.

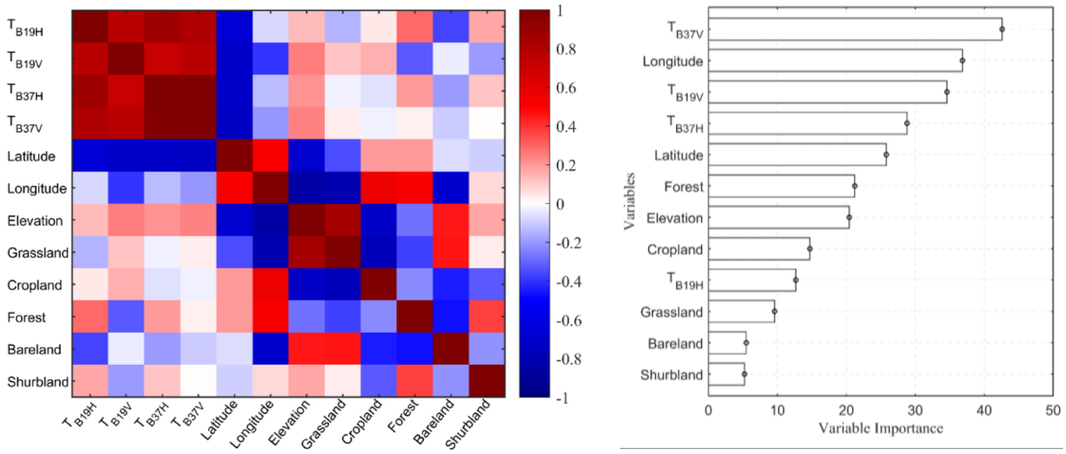

Figure 3. Correlations between the predictor variables (left) and the ranking of variable importance (right). The importance of variables, referred to as Mean Decrease Accuracy (MDA) in RF model, is obtained by averaging the difference in out-of-bag error estimation before and after the permutation over 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 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 three tests but with different validation datasets. In Test1, the validation data are from out-of-bag (OOB) samples. Generally, in the RF model, approximately two-thirds of the samples (in-bag samples) are used to train the trees and the remaining one-third (OOB samples) are used to estimate how well the fitted RF algorithm performs. This preliminary assessment offers 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 (Fig. 3). 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

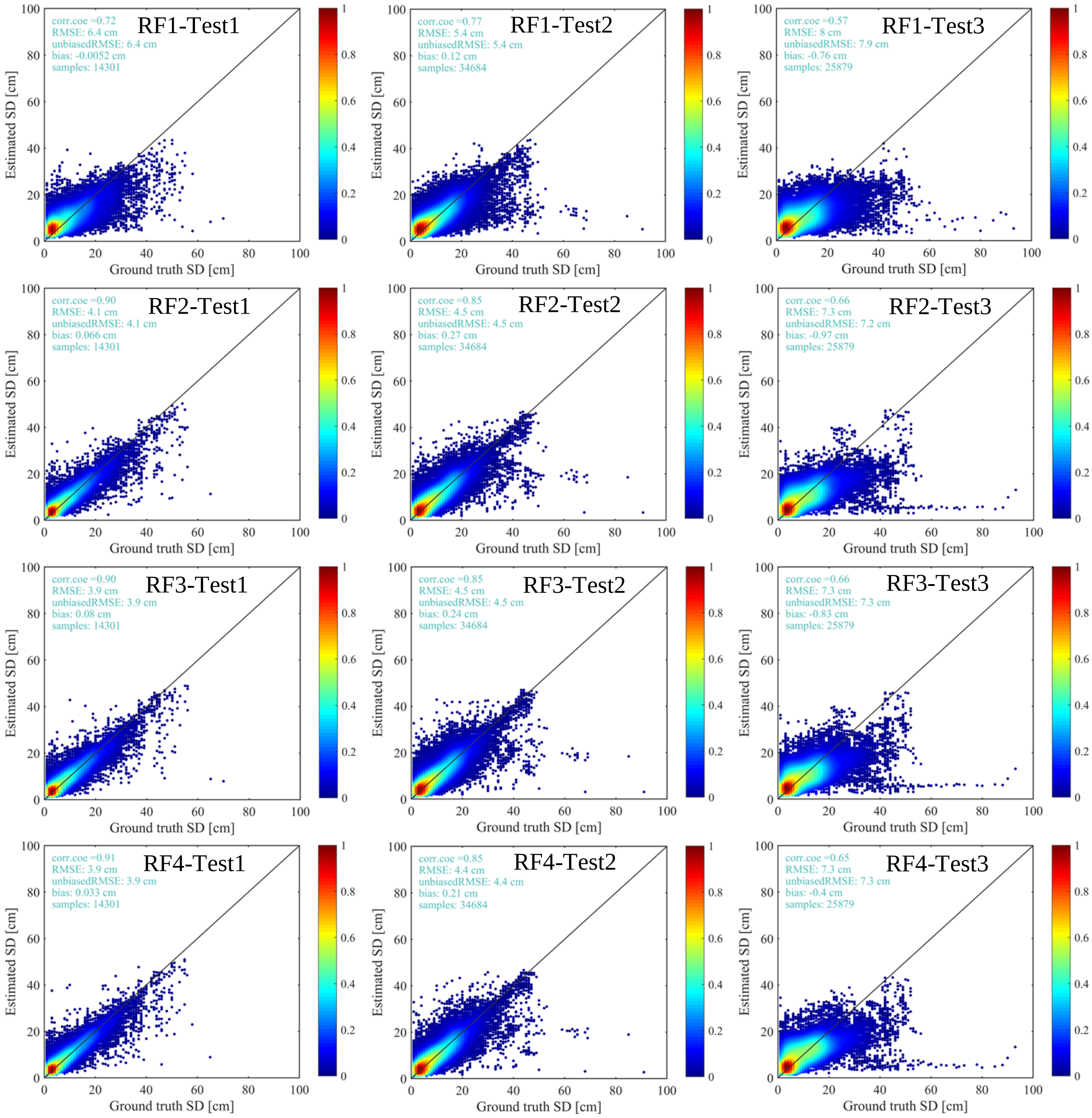

Figure 4. The color-density scatterplots of the estimated snow depth with four fitted RF algorithms and the ground truth snow depth. The four trained RF algorithms (RF1, RF2, RF3, RF4) were evaluated with three validation datasets (Test1, Test2, Test3).

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. The 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).

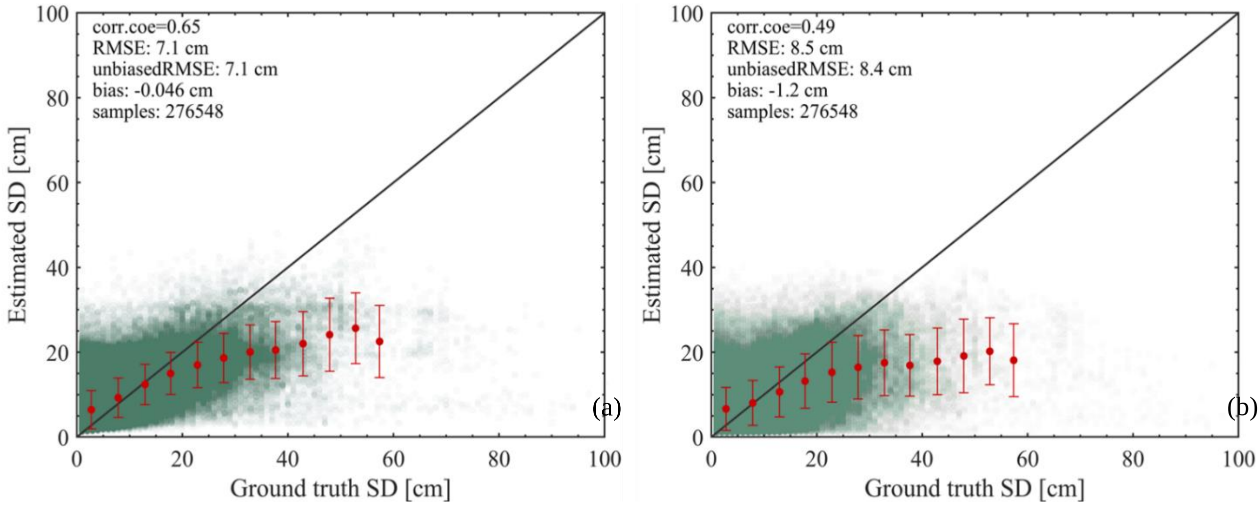

Figure 5. Scatterplots of the estimated snow depth and the ground truth observation for (a) RF and (b) WESTDC products.

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.

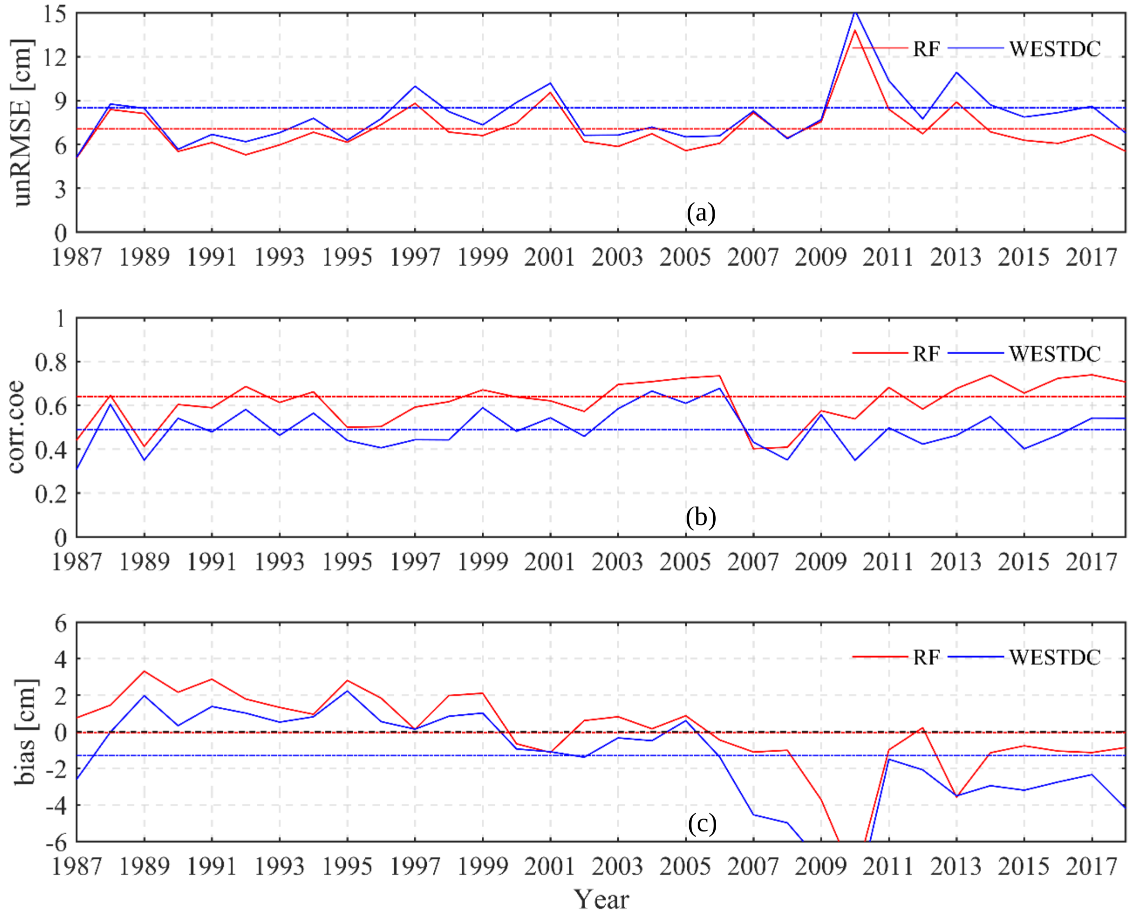

Figure 6. Time series of (a) unbiased RMSE (unRMSE), (b) correlation coefficient (corr.coe) and (c) bias for RF and WESTDC products. The colorful dashed lines represent mean values of assessment indexes.

The assessment of snow depth product was performed in three snow cover areas in China. We selected 20 cm as a threshold to assess the performances in deep (> 20 cm) and shallow (≤ 20 cm) snow cover. Table 3 displays the comparison between RF estimates and WESTDC product in the three snow cover areas. Both products present notable underestimation of deep snow cover, with the biases of -34.1 cm and -33.8 cm in QTP for the RF and WESTDC products, respectively. The biases are -10.4 cm and -8.9 cm in NE and northern XJ for RF product, respectively, whereas they are -11.8 cm and -13.2 cm for WESTDC data. For shallow snow cover, the RF product is superior to the WESTDC estimates in QTP, with unbiased RMSEs of 3.4 cm (RF) and 5.6 cm (WESTDC). Furthermore, the WESTDC product presents an overestimation in QTP, with a bias of 4.0 cm that is much higher than RF's 0.6 cm. The unbiased RMSEs of the RF product are 5.4 cm and 6.1 cm in NE and northern XJ for shallow snow cover, respectively, lower than the WESTDC's values of 6.5 cm and 7.4 cm.

In the Discussion, we list the potential errors of the reconstrued snow depth (Page 10, Line 18-28 and Page 11, Line 1-13, in the revised manuscript).

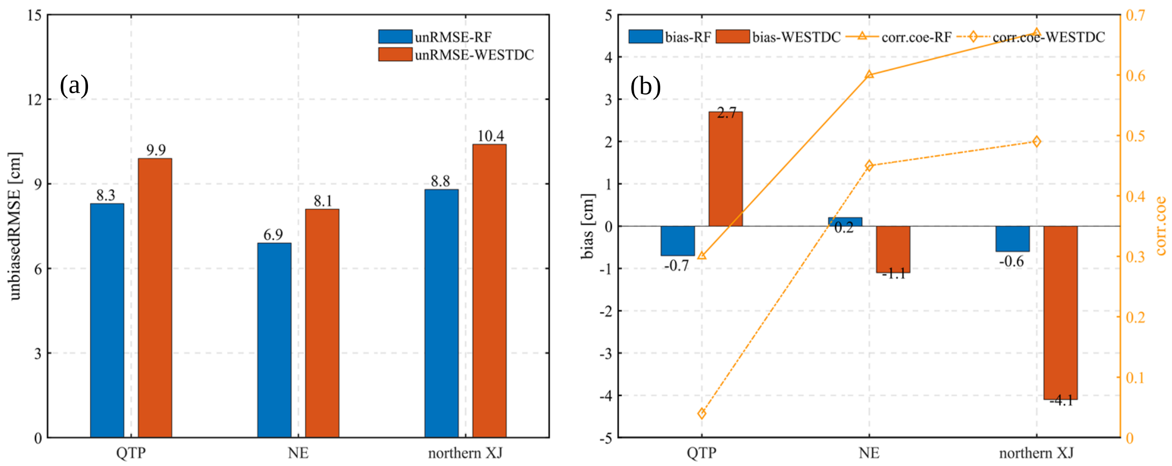

Figure 7. The validation of RF and WESTDC snow depth products in the three stable snow cover areas in China with respect to (a) the unbiased RMSE, (b) bias and correlation coefficient.

Table 3. Comparison between RF estimates and WESTDC product in three stable snow cover areas for deep (> 20 cm) and shallow (≤ 20 cm) snow cover.

RF product						
Regions	QTP		NE		northern XJ	
SnowDepth (cm)	<= 20	> 20	<= 20	> 20	<= 20	> 20
corr.coe	0.30	0.06	0.49	0.17	0.48	0.31
bias (cm)	0.59	-34.12	1.79	-10.38	2.52	-8.85
unRMSE (cm)	3.43	20.70	5.36	7.00	6.12	9.62
Samples	15503 (96.4%)	583 (3.6%)	151939 (87.3%)	22168 (12.7%)	32468 (69.8%)	14051 (30.2%)
WESTDC product						
Regions	QTP		NE		northern XJ	
SnowDepth (cm)	<= 20	> 20	<= 20	> 20	<= 20	> 20
corr.coe	0.16	-0.18	0.37	0.03	0.34	0.16
bias (cm)	4.02	-33.78	0.47	-11.75	-0.39	-13.22
unRMSE (cm)	5.60	21.62	6.47	9.10	7.35	11.30
Samples	15503 (96.4%)	583 (3.6%)	151939 (87.3%)	22168 (12.7%)	32468 (69.8%)	14051 (30.2%)

Major issues:

1. I agree with the Anonymous Reviewer (point 9b), who pointed that this complicated methodology of using RF to produce more data is probably unnecessary and that it should be tested whether this step is necessary and whether it does increase uncertainty to the product or not.

Response 1: Thank you for your comments. We tested four RF algorithms trained with different predictor variables (Fig. 4). The results showed that the accuracy of RF model is greatly influenced by geographic location, elevation, and land cover fractions. However, we also found redundant predictor variables due to high correlation. The elevation variable is highly correlated (correlations higher than 0.9) with geographic location (Fig. 3). Additionally, the correlation between longitude and land cover type (e.g., grassland, cropland, forest and bareland) is significant. Thus, land cover type and elevation are not necessary. We directly used a simple RF algorithm to retrieve a consistent 32-year daily snow depth dataset from 1987 to 2018. Please refer to the response to "General Comments" above.

2. The most important issue is that the validation of the RF and the pixel based snow depths is not fair. This is because stations used for validation are only a temporal subsample of the training station set. The spatial sub-sampling was not conducted, i.e., stations for all geographic locations were used for training and validation. This is a very important problem, because latitude and longitude are the third and fourth important predictors in the model, nearly as important as the Tb. The RF model cannot know values of this predictors already during training, because the validation does not make sense. Therefore, the errors reported in this study are very optimistic (underestimated) and should be recalculated using 50% of the stations (not the data, i.e. spatial not temporal subset) which were not used to train the RF model. The same 50% subset should be used to validate the pixel-based method.

Response 2: Thank you for your constructive comments. Other reviewers gave similar comments. 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. Please refer to the response to "General Comments" above.

3. The pixel based SD product effectively fails to model snow above 20cm depth (Figure9). This is a serious limitation and it should be explained very deeply in the discussion: (1) why this happens, (2) what is the true applicability of the product given the RMSE is 5cm.

Response 3: Thank you for your comments. We discussed this in Section 4.3.

"Fig. 7 indicates that the RF model does not fully solve the overestimation and underestimation problems. For deep snow (> 20 cm), the biases are up to -8.9 cm and -10.4 cm in NE and northern XJ, respectively. Deep snow conditions account for roughly 10% of all training samples (Fig. 2). The estimates for deep snow cover in the QTP exhibit a large bias of -34.1 mm. Fig. 6 also illustrates that the fitted RF algorithms have no predictive ability for extremely deep snow conditions, especially in QTP. We checked the training data and found that the extreme high snow depth data (> 60 cm) occurred in QTP. However, the number of such cases is very small. In addition, the station measurements are point values while the satellite grids have a spatial resolution of 25 km × 25 km. Thus, the representativeness of these data is questionable. Snow depth estimation in the mountains remains a challenge (Lettenmaier et al., 2015; Dozier et al., 2016; Dahri et al., 2018). Numerous studies have been conducted on the snow cover over the QTP and have indicated that the snow cover in the Himalayas is higher than elsewhere, ranging from 80% to 100% during the winter (Basang et al., 2017; Hao et al., 2018). Additionally, Dai et al. (2018) showed that deep snow (greater than 20 cm) was mainly distributed in the Himalayas, Pamir, and Southeastern Mountains. Thus, the RF product produced in this paper has poor performance in QTP for deep snow cover."

4. The methods are very difficult to follow, I noticed that the other Reviewers managed to understand them better than me, but still, I am not completely sure how the study was conducted. This entire chapter should be rewritten, simplified and better structured. Often different words are used in the same context to name the same things, what makes understanding of this paper even more difficult (see attachment for some examples). The results and discussion sections are very poorly written: methods, results and discussion are mixed in each of these sections (see attachment for some examples).

Response 4: Thank you for your comments. We revised the manuscript carefully and thoroughly to make paper structure clearer. Additionally, a thorough revision of the manuscript was completed by a native speaker.

5. Authors should also justify better why this is a real-time approach. Is there an operational implementation of this algorithm?

Response 5: We removed the word 'real-time' in the revised manuscript.

6. Eventually, Authors claim that ML in RS is a very novel research problem, e.g. "Machine learning (ML) is a common method used in many research fields, and its early application in remote sensing is promising". The applications of ML in RS are not early, they are in RS since decades, either for regression (as in this study) or classification. The use of RF for regression, cannot be understand as a novelty, because it simply is not. Authors should better explain in which aspects the MS is novel.

Response 6: We apologize for the ambiguous description. We rewrote this paragraph as follows: "Over the last two decades, RF has been one of the most successful ML algorithms for practical applications due to its proven accuracy, stability, speed of processing and ease of use (Rodriguez-Galiano et al., 2012; Belgiu et al., 2016; Maxwell et al., 2018; Bair et al., 2018; Qu et al., 2019; Reichstein et al., 2019, Tyralis et sl., 2019a)" (Page 3, Line 2-5, in the revised manuscript).

In Section 2.2, we listed some advantages of the RF model. (Page 4, Line 19-30, in the revised manuscript).

We agree with your opinion that machine learning method is not novel in remote sensing and have rewritten the sentence. It now reads, "The primary objectives of this study are to assess the feasibility of the RF model in estimating snow depth, to determine whether the inclusion of auxiliary information (geolocation, elevation and land cover fraction) contributes to the improvement of RF, and eventually to develop a time series (1987 to 2018) of snow depth data in China and analyze the trends in annual mean snow depth. To complete the feasibility study of the RF model, we designed four RF algorithms trained with different combinations of predictor variables and validated them using independent reference data temporally and spatially. To our knowledge, this type of assessment of RF algorithm performance has not been made to date over China" (Page 3, Line 7-12, in the revised manuscript).

Minor issues: (from hand-written comments)

1. Page 1, line 20, the applications of ML in RS are not early, please remove early.

Response 1: Word ‘early’ removed.

2. Page 1, Line 22, from 1987-2018.

Response 2: We changed the sentence to ‘during the period 1987-2018.’

3. Page 1, Line 23, ‘the advanced microwave scanning radiometer’. The first letter should be capitalized.

Response 3: We selected SSM/I and SSMIS data as satellite observations and thus deleted this description.

4. Page 2, Line 23, this paper is about snow depth, not SWE.

Response 4: Thank you for your comments. We rewrote it as “Snow depth is a crucial parameter for climate studies, hydrological applications and weather forecasts (Foster et al., 2011; Takala et al., 2017; Tedesco et al., 2016; Safavi et al., 2017).”

5. Page 4, Line 8, not prediction, but regression.

Response 5: We changed ‘prediction’ to ‘regression.’

6. Page 4, Line 24, 25*25km^2 ? ?

Response 6: It is 25 km x 25 km. We selected SSM/I and SSMIS data as satellite observations to retrieve snow depth in the revised manuscript and thus deleted this description.

7. Page 6, Line 7, cold overpass ? ?

Response 7: Thank you for your comments. We rewrote this sentence as ‘To avoid the influence of wet snow, only ascending (F08) and descending (F11, F13 and F17) overpass data were used (Table 1).’

Table 1. 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

8. Page 6, Line 13-15, station data is daily? What is harsh climate?

Response 8: Thank you for your comments. We rewrote these sentences as ‘The weather station daily data in China from 1987 to 2018 were provided by the National Meteorological Information Centre, China Meteorology Administration (CMA, http://data.cma.cn/en)’ and

‘The sites are not distributed homogeneously, and few are located in inaccessible regions with extreme climates and complex terrain conditions, e.g., the western part of QTP.’

9. Page 6, Line 22, snow depth can be over 70 cm

Response 9: Thank you for your comments. Fig. 2 showed the histograms of snow depth observations from training and testing stations. 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 therefore not evident. In the revised manuscript, we maintained these data.

10. Page 7, Line 15-19, the description is not clear.

Response 10: Thank you for your comments. We rewrote this paragraph in Section 2.2.1.
‘2.2.1 Random forest
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)."

11. Page 7, Line 27-28, why you asking questions here. Page 8, Line 3, 80000 pairs? Not clear

Response 11: We deleted the questions and rewrote this paragraph in Section 2.2.2.
‘(2) Training sample size
One of the advantages of the RF model is that it can effectively handle small sample sizes (Biau and Scornet et al., 2016). A test was conducted to demonstrate the insensitivity of the RF model to the training sample size. The input predictor variables include geographic location and T_B (Table 2, RF2). The flowchart of the test process is shown in Fig. 4. To ensure a sufficient number of samples, 80,000 records from 1987 to 2004 were used to test the required size of the training samples and a two-year stand-alone dataset from (2005-2006) was applied to assess the performance. During this process, the number of

samples selected randomly was from 5000 to 80,000 (step, 5000). We consider three evaluating indicators (the unbiased root mean square error (RMSE), bias and correlation coefficient) to illustrate the stability of the RF model to the training sample size."

12. Page 8, Line 4-8, what is this paragraph about? What is 'stability'? in respect to what?

Response 12: Thank you for your comments. We tested the sensitivity of the RF model to the training sample size. We rewrote this paragraph. Please refer to the response to "Minor Comment 11" above.

13. Page 8, Line 15-26, this is ambiguous. Which radiation is scattered by snow? Which radiation the snow is transparent? What is the snow of these radiations? Perhaps some of the radiation is radiated by snow itself, not scattered…

Response 13: Thank you for your comments. Most passive microwave snow depth retrieval algorithms exploit the negative spectral gradient between measurements at 19 GHz and 37 GHz. We rewrote this paragraph in Section 2.2.2.

'All available channels on the SSM/I and SSMIS are listed in Table 1. The 23 GHz channel is sensitive to water vapor and not surface scattering, which introduces uncertainty to the estimation process (Ji et al., 2017). The 85 (91) GHz channel is seriously influenced by the atmosphere (Kelly et al., 2009; Xue et al., 2017). Typically, the lower frequency (19 GHz) is used to provide a background T_B against which the higher frequency (37 GHz) scattering-sensitive channels are used to retrieve snow depth.'

14. Page 9, Line 2-4, this sentence should move to the introduction section.

Response 14: We left out the pixel-based method in this paper due to RF's limitations.

15. Page 9, Line 6-7, 19GHz is always 18GHz.

Response 15: Thank you for your comments. We used the same symbol in the manuscript. 'In this study, the difference between 19.35 (36.5) GHz and 18.7 (37) GHz was ignored (hereafter referred as 19 GHz and 37 GHz, respectively).'

16. Page 9, Line 24-25, seasons, should be season or months. Isn't wet snow likely in November?

Response 16: We changed the word 'seasons' to 'season.' Although a snow cover detection method (Grody et al., 1996) was used to filter out wet snow conditions, wet snow is still possible in November.

17. Page 10, Line 1-3, some repetition, not clear.

Response 17: We modified the sentence to "The sensitivity of the RF model to the training sample size was conducted to confirm the appropriate number of training samples."

18. Page 10, Line 5, the term 'represents' is changed to 'presents'. RMSE ranges.., not RMSEs range…

Response 18: Thank you for your comments. We rewrote this sentence as ‘Fig. 4a presents unbiased RMSE ranges from 5.1 cm to 5.5 cm.’

19. Page 10, Line 10, what is the optimal number you chosen here?

Response 19: According to the sensitivity analysis, the number of training samples has less influence on the prediction accuracy of the RF model. In our study, we selected all available samples (28602) from training stations (Fig. 1) during the period 2012-2014 to train the RF models.

20. Page 10, Line 11, this statement is not supported by the results.

Response 20: One of the advantages of the RF model is that it can effectively handle small sample sizes (Biau and Scornet et al., 2016). Our results also indicated that the performance of RF model is insensitive to the training sample size.

21. Page 10, Line 16-18, please move this sentence to the method section.

Response 21: We moved it to Section 2.2.2.

22. Page 10, Line 23, this is discussion, not result.

Response 22: It was moved to Section 4.3.

23. Page 11, Line 2-3, how the relative error was calculated?

Response 23: $RPE=abs(bias*100/SD_{ground})$.

24. Page 11, Line 6-8, is method, not result.

Response 24: Moved.

25. Page 11, Line 11-13, the reference?

Response 25: We added the reference and moved this sentence to the discussion section. " Second, the large diurnal temperature range tends to subject the snowpack to frequent freeze-thaw cycles and leads to rapid snow grain (~2 mm) and snow density (200-350 kg/m^3) growth and consequently a high TB difference (Meløysund et al., 2007; Durand et al., 2008; Yang et al., 2015; Dai et al., 2017).”

26. Page 11, Line 16-19, aren’t only cold/night orbits data used?

Response 26: In this study, a snow cover detection method is used to filter out wet snow cover; however, there are still misclassification errors, especially at the end of winter (Liu et al., 2018).

Liu, X., Jiang, L., Wu, S., Hao, S., Wang, G., and Yang, J.: Assessment of Methods for Passive Microwave Snow Cover Mapping Using FY-3C/MWRI Data in China, Remote Sensing, 10, 524-539, 10.3390/rs10040524, 2018.

27. Page 11, Line 25-30, this is how to judge base on the maps?

Response 27: We moved this sentence to the discussion.

28. Page 12, Line 12-16, mixing results and discussion!

Response 28: We moved this sentence to Section 4.3.

29. Page 12, Line 25-27, move to method section!

Response 29: In the revised manuscript, we left out the pixel-based method and thus deleted this sentence.

30. Page 13, Line 3-4, where and why?

Response 30: We deleted this sentence because the pixel-based method was left out in the revised manuscript.

31. Page 13, Line 12-13, this sentence should be "To evaluate the long-term…."

Response 31: We corrected this sentence.

32. Page 13, Line 23, where is the comparison?

Response 32: We rewrote it as "The overall accuracy of the RF product is higher than the WESTDC estimates, with unbiased RMSEs of 7.1 cm and 8.5 cm, respectively (Fig. 7a and 7b)."

33. Page 15, Line 3-13, move to results section.

Response 33: Done.

34. Page 15, Line 17-22, only cold/night orbits data were used in winter season, how to explain it?

Response 34: Please refer to the response to "Minor Comment 26" above.

35. Page 15, Line 22, It is result, not discussion.

Response 35: Moved.

36. Page 16, Line 2, "H-pol" is "in horizontal polariton".

Response 36: Corrected.

37. Page 16, Line 8-15, not clear explanation. Not 'predictor importance' but 'predictor variable importance'.

Response 37: We modified the sentence to "The importance of predictor 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" (Page 19, Line 6-9, in the revised manuscript).

38. Page 16, Line 12, remove the 'by far', 'more dependent on station data' is changed to 'geographically dependent'.

Response 38: Done.

39. Page 16, Line 17-27, the result does not support this because DEM was not a predictor variable in this paper. If DEM is better than lat/lon, why not use DEM?

Response 39: We redesigned the procedure and included the DEM as one of the predictor variables (Table 1). Fig. 3 indicates that DEM is highly correlated with the geolocation (lat/lon).

40. Page 17, Line 2, Significantly? Statistical test conducted?

Response 40: It means that there is a notable accuracy difference for different land cover types. We deleted the word 'significantly.'

41. Page 17, Line 3, what if land cover changes over time?

Response 41: This is a wonderful question. In this study, we assume the land cover type does not change. We can study this in future work.

42. Page 17, Line 15-29, These sentences belong to method section.

Response 42: The aim of this part is to demonstrate that more prior snow information can improve the performance of the RF model. According to Reviewer #4's suggestion, we omitted this and will present it in a future publication.

43. Page 18, Line 4-6, This part is discussion.

Response 43: We moved it to Section 4.1.

44. Page 18, Line 8, where is this method?

Response 44: The method is the pixel-based algorithm. We omitted this part.

45. Page 18, Line 11, past or present

Response 45: We revised the manuscript carefully and thoroughly to make the tense correct.

46. Page 18, Line 15, than the former…

Response 46: word 'former' added.

47. Page 18, Line 16-20, is this really a conclusion? Page 18, Line 21, This is not a conclusion, but summary. What is the conclusion here? I do not find…

Response 47: We rewrote the conclusion (Page 11, Line 14-28, Page 12, Line 1-16, in the revised manuscript).
"The present study analyzed the application of the RF model to snow depth estimation at temporal and spatial scales. Temporally and spatially independent datasets were applied

to verify the fitted RF algorithms. The results suggested that the accuracy of fitted RF algorithms was greatly influenced by auxiliary data, especially the geographic location. However, the inclusion of strongly correlated predictor variables (elevation and land cover fraction) did not further improve the RF estimates. Therefore, in some cases, a few representative predictor variables should be selected. Due to naive extrapolation outside the training range, the transferability of a fitted RF algorithm at the temporal scale was better that that in spatial terms, e.g., with unbiased RMSEs of 4.5 cm and 7.2 cm for the RF2 algorithm, respectively.

In this study, the fitted RF2 algorithm was used to retrieve a consistent 32-year daily snow depth dataset from 1987 to 2018. Then, an evaluation was carried out using independent reference data from the validation stations during the period 1987-2018. The overall unbiased RMSE and bias were 7.1 cm and -0.05 cm, respectively, outperforming the WESTDC product (8.4 cm and -1.20 cm). In QTP, the unbiased RMSE and bias of RF estimates for shallow (≤ 20 cm) snow cover were 3.4 cm and 0.59 cm, respectively, much lower than WESTDC's 5.6 cm and 4.02 cm. In NE and northern XJ, RF estimates were superior to the WESTDC product but still presented an underestimation for deep snow (> 20 cm), with biases of -10.4 cm and -8.9 cm, respectively.

Three long-term (1987-2018) datasets, including ground truth observations, RF estimates and WESTDC product, were applied to analyze the trends of snow depth variation in China. The results suggested that there existed different trends among the three datasets. The overall trend of snow depth in China presented a significant increasing based on the ground truth observations, with a correlation coefficient of 0.57. Moreover, the trend in NE was highly consistent with the overall trend in China, with a correlation coefficient of 0.64. Neither the WESTDC nor the RF product presented significant trends except in QTP. The WESTDC product showed a significant decreasing trend in QTP, with a correlation coefficient of -0.55, whereas there were no significant trends for ground truth observations and the RF product.

As discussed in Section 4, our reconstructed snow depth estimates are still challenged by several problems, e.g., underestimation for deep snow. Additional prior knowledge of snow cover, such as snow cover fraction, snow density, and snow grain size, is necessary to improve the RF model. Combining the snow forward model with the ML method will be the focus of future work. Furthermore, the mass balance approaches, e.g., the Parallel Energy Balance model, will be used to improve the snow depth retrievals in high-altitude areas. In addition, although our results indicate that the RF method is a promising potential tool for snow depth estimation, there are a few pitfalls such as the risk of naive extrapolation and poor transferability in spatial terms limiting its application in spatio-temporal dynamics. It is in addressing these shortcomings that the techniques of deep learning promise breakthroughs. We are attempting to operate the Deep Neural Networks (DNN) model to overcome the limitations of traditional ML approaches."

Response to Reviewer Comments by Nir Krakauer 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 #4

General Comments: The basic theme of this manuscript, the application of random forest (RF) to provide an empirical transfer model from remotely sensed radiances to snow depth, has merit, given that physically based transfer models are subject to limitations. However, some of the modeling choices appear questionable and should be better justified or simplified. The RF modeling described in Section 2.3 has the following main components: (1) Using SSMI data from 1987-2004 for training and from 2005-2006 for validation, in order to evaluate the number of training samples required for good accuracy. (2) Using AMSR2 data from 2014-2015 for training and from 2012-2013 for validation. Snow depth estimated by this model is then used to generate an approximate spatially varying relationship between 2 SSMI channel radiances and snow depth. The resulting simple SSMI-based formula is used to reconstruct estimated snow depth for 1987-2018, which is validated for 2017-2018.

Specific comments:

1. Approach (2) appears unnecessarily complicated. If the goal is to establish a product for 1987-2018, where only SSMI inputs are available for the entire period, it is more logical to train an RF model directly with SSMI inputs (as done in (1) – not with AMSR2 inputs) fitted to station data (not reconstructed data). If the authors want to retain their more complicated approach, they should compare it to the simpler one to demonstrate that it actually has superior accuracy.

Response 1: We agree with the reviewer's opinion, and these suggestions are very constructive. Other reviewers gave us similar comments. Thus, we directly selected SSM/I and SSMIS data as satellite observations in the revised manuscript.

The procedure described in the original manuscript was complicated. Based on the correlations between the predictor variables and the variable importance metrics (Fig. 1), 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.

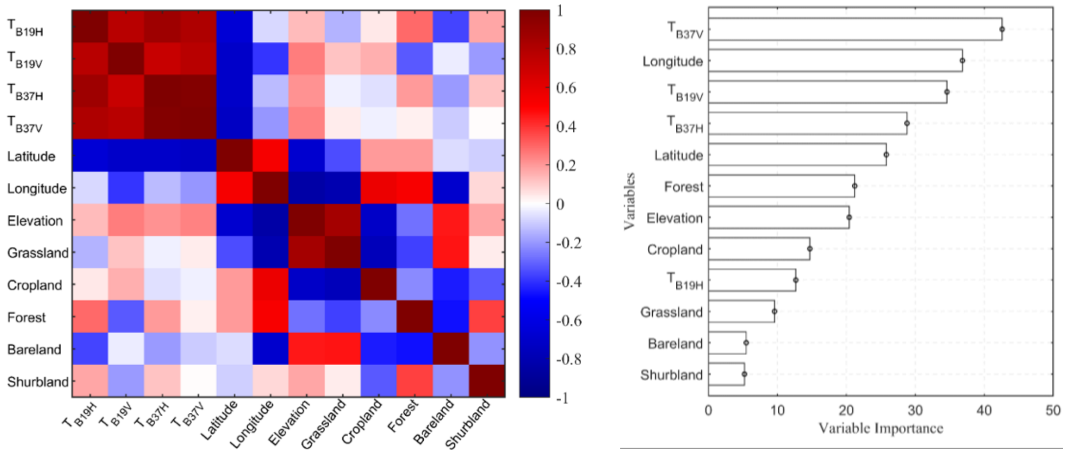

Figure 1. Correlations between the predictor variables (left) and the ranking of variable importance (right). The importance of variables, referred to as Mean Decrease Accuracy (MDA) in RF model, is obtained by averaging the difference in out-of-bag error estimation before and after the permutation over 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 training sample.

<table>
<tr><th>Name</th><th>Predictor Variables</th><th>Target</th><th>Note</th></tr>
<tr><td>RF1</td><td>T_{B19V}, T_{B37V}</td><td rowspan="4">snow depth</td><td rowspan="4">land cover types: grassland, cropland, bareland, shurbland, forest</td></tr>
<tr><td>RF2</td><td>T_{B19V}, T_{B37V}, Latitude, Longitude</td></tr>
<tr><td>RF3</td><td>T_{B19V}, T_{B37V}, Latitude, Longitude, Elevation</td></tr>
<tr><td>RF4</td><td>T_{B19V}, T_{B37V}, Latitude, Longitude, Elevation, Land cover fraction</td></tr>
</table>

Then, we conducted three tests to verify the fitted RF algorithms (Table 1). The same training samples (same algorithms) were used for the three tests but with different validation datasets. In Test1, the validation data are from out-of-bag (OOB) samples. Generally, in the RF model, approximately two-thirds of the samples (in-bag samples) are used to train the trees and the remaining one-third (OOB samples) are used to estimate how well the fitted RF algorithm performs. This preliminary assessment offers 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 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. 2 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 to 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

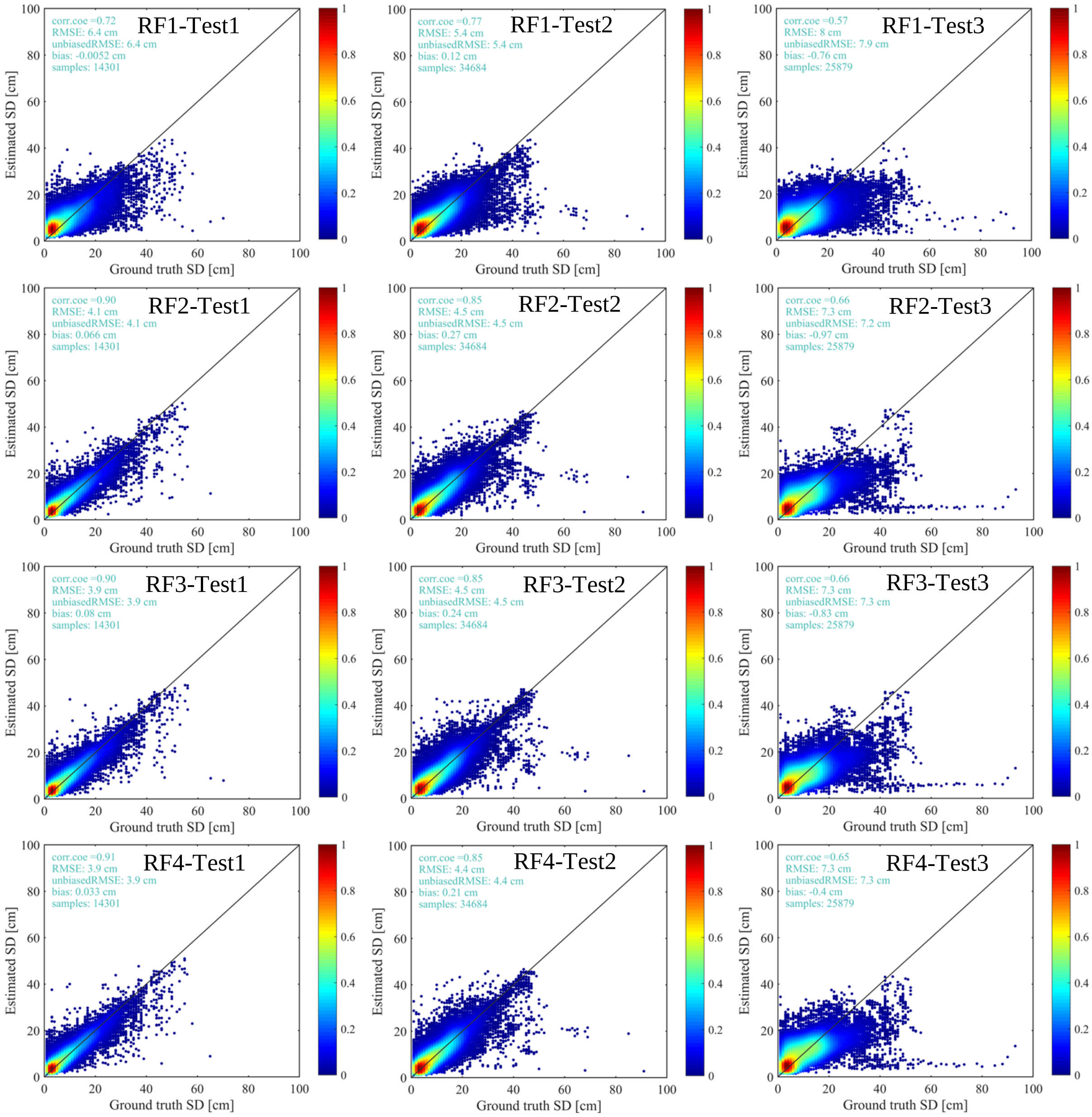

Figure 2. The color-density scatterplots of the estimated snow depth with four fitted RF algorithms and the ground truth snow depth. The four trained RF algorithms (RF1, RF2, RF3, RF4) were evaluated with three validation datasets (Test1, Test2, Test3).

Finally, we directly used the fitted RF2 algorithm to retrieve a consistent 32-year daily snow depth dataset. It was evaluated against the independent ground truth measurements from the validation stations (Fig. 6) 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).

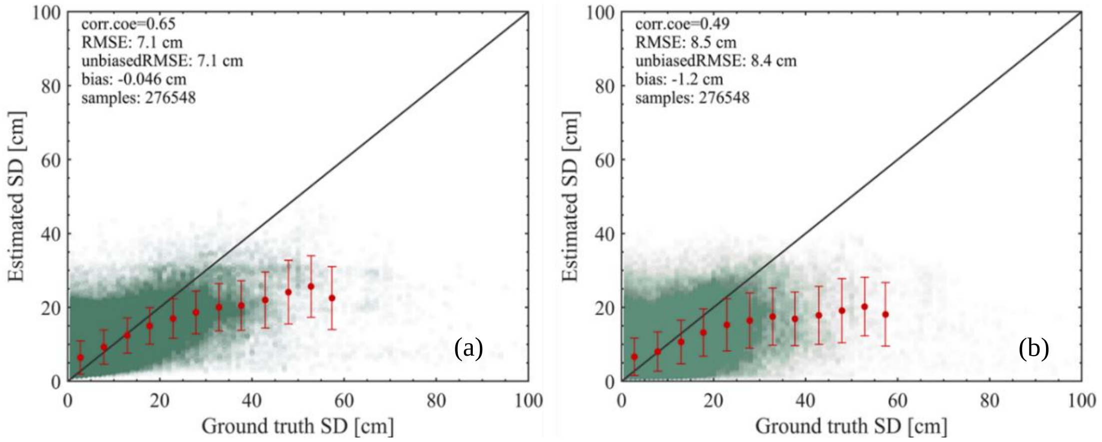

Figure 3. Scatterplots of the estimated snow depth and the ground truth observation for (a) RF and (b) WESTDC products.

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. 4. The results show that the RF estimates outperform the WESTDC product with respect to unbiased RMSE and correlation coefficient from season to season.

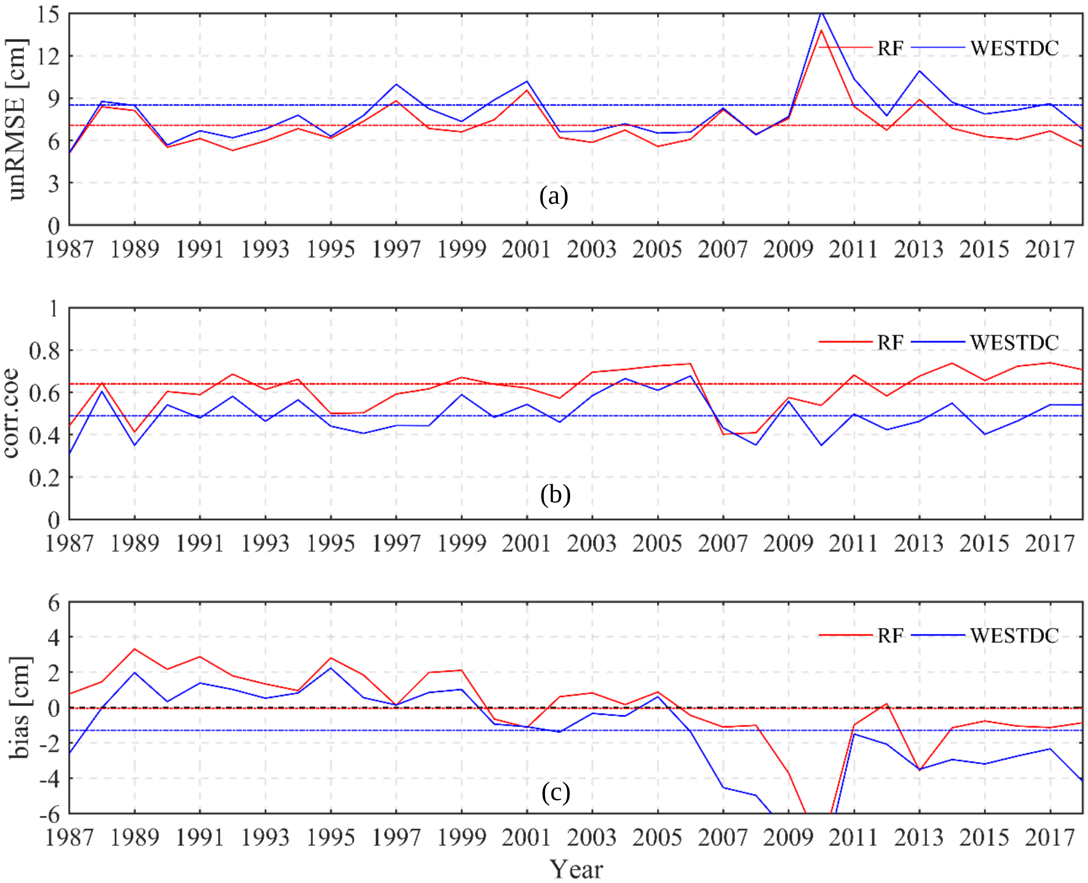

Figure 4. Time series of (a) unbiased RMSE (unRMSE), (b) correlation coefficient (corr.coe) and (c) bias for RF and WESTDC products. The colorful dashed lines represent mean values of assessment indexes.

The assessment of snow depth product was also performed in three snow cover areas in China for shallow (≤ 20 cm) and deep snow cover (> 20 cm).

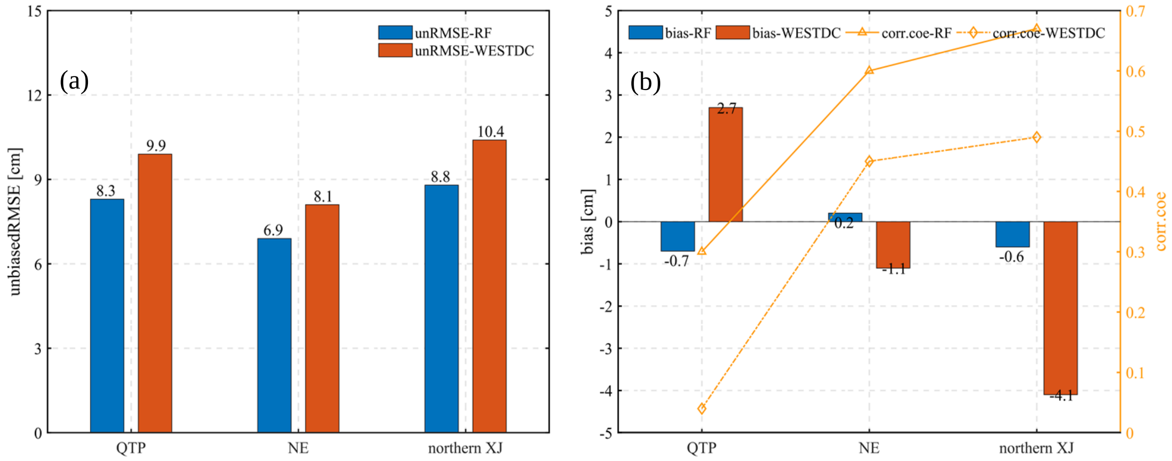

Figure 5. The validation of RF and WESTDC snow depth products in three stable snow cover areas in China with respect to (a) the unbiased RMSE, (b) bias and correlation coefficient.

Table 3. Comparison between RF estimates and WESTDC product in three stable snow cover areas for deep (> 20 cm) and shallow (≤ 20 cm) snow cover.

RF product						
Regions	QTP		NE		northern XJ	
SnowDepth (cm)	<= 20	> 20	<= 20	> 20	<= 20	> 20
corr.coe	0.30	0.06	0.49	0.17	0.48	0.31
bias (cm)	0.59	-34.12	1.79	-10.38	2.52	-8.85
unRMSE (cm)	3.43	20.70	5.36	7.00	6.12	9.62
Samples	15503 (96.4%)	583 (3.6%)	151939 (87.3%)	22168 (12.7%)	32468 (69.8%)	14051 (30.2%)
WESTDC product						
Regions	QTP		NE		northern XJ	
SnowDepth (cm)	<= 20	> 20	<= 20	> 20	<= 20	> 20
corr.coe	0.16	-0.18	0.37	0.03	0.34	0.16
bias (cm)	4.02	-33.78	0.47	-11.75	-0.39	-13.22
unRMSE (cm)	5.60	21.62	6.47	9.10	7.35	11.30
Samples	15503 (96.4%)	583 (3.6%)	151939 (87.3%)	22168 (12.7%)	32468 (69.8%)	14051 (30.2%)

2. There is another way to tackle the problem of different microwave satellite sensors being available over different portions of the 1987-2018 period, which the authors may also want to consider. This would involve combining estimates from multiple fitted RF models, one for each satellite sensor available for part of the time period, which would potentially more fully use the partly-independent information from multiple satellite sources, which may each have different wavelength ranges, overpass times, and other sensor characteristics.

Response 2: These suggestions are very constructive. However, as a change from the original manuscript, we resorted to using only SSM/I and SSMIS data as satellite observations in this study. As shown in Table 4 below, the characteristics of these sensors are sufficiently similar to assume that an algorithm defined for one sensor can be applicable to the next. We have rewritten the introduction of satellite data in Section 2.1: "The SSM/I and SSMIS sensors are suitable for producing a long-term consistent snow depth dataset due to their similar configurations and intersensor calibrations (Armstrong et al., 1994)" (Page 3, Line 21-23, in the revised manuscript).

Table 4. 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. Another issue is the training/validation station data split. As one of the other reviewers points out, in order to better estimate the error at ungauged sites, it makes more sense to not use some stations at all for training and retain them for validation, instead of validating with data for the same stations but different years.

Response 3: Thank you for your comments. One of the major issues of this study is that the validation data are not temporally and spatially independent. Thus, available stations in China were randomly divided into two roughly equal-sized parts by Matlab software (Fig. 6). 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. 7 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. 7.

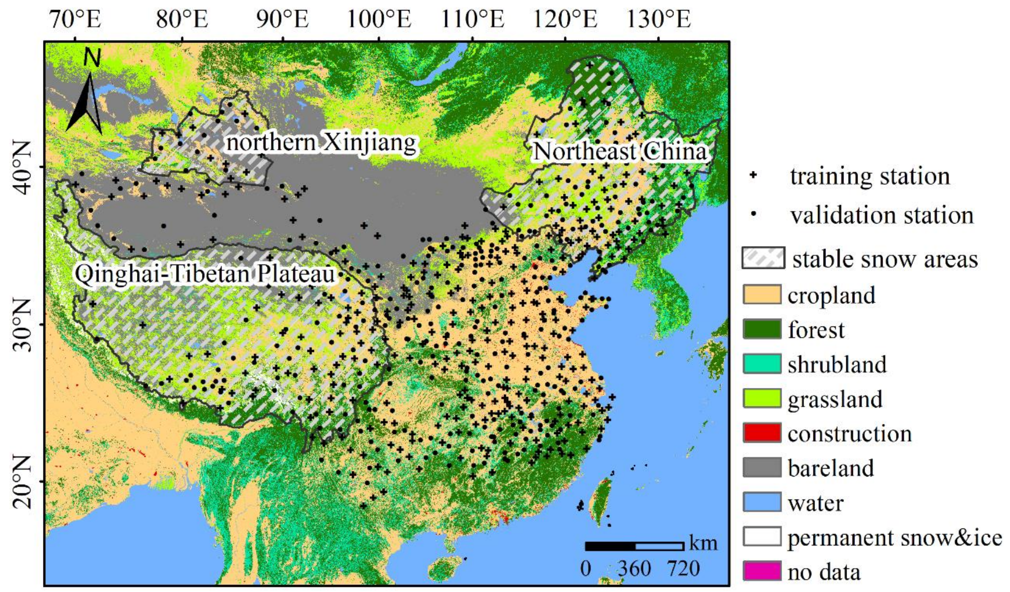

Figure 6. Spatial distribution of the weather stations and land cover types in the study area. There are three stable snow cover areas in China: Northeast China (NE), northern Xinjiang (XJ) and the Qinghai-Tibetan Plateau (QTP).

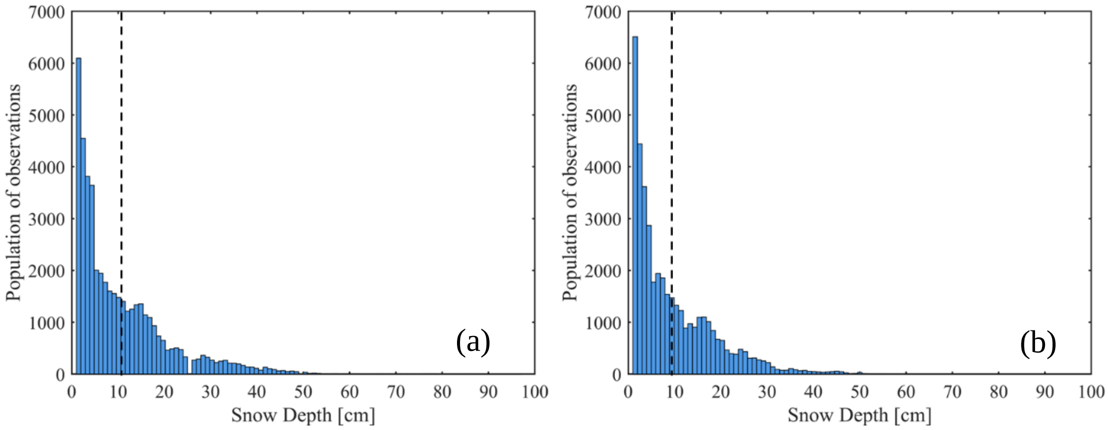

Figure 7. Histograms of snow depth observations from (a) training and (b) validation stations. The average values (black dashed lines) are equal to 10.5 cm and 9.8 cm, respectively.

4. There is no comparison presented between the RF method and physically based transfer models or existing satellite or reanalysis snow products over China. This work would be stronger if the authors can conduct such a comparison and show whether RF in fact leads to improvements in snow estimation beyond existing approaches.

Response 4: Thank you for your comments. The linear-fitting method was developed based on SSM/I observations and station snow depth data by 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). Yang et al. (2019) demonstrated that the WESTDC product outperforms four other snow depth datasets in China. Thus, in this study, we directly compared the RF estimates with the WESTDC product.

We also show that an overall improvement of 15.4 % in China is achieved compared to the WESTDC product (Fig. 3). In QTP, the unbiased RMSE and bias of RF estimates for shallow (≤ 20 cm) snow cover were 3.4 cm and 0.59 cm, respectively, much lower than WESTDC's 5.6 cm and 4.02 cm (Table 3). Please refer to the response to "Specific comment 1" above.

[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.

5. Section 4.5 discusses the performance of an RF model under an ensemble of simulated weather conditions and microwave radiances. It is not clear what this section adds to the stronger results of the earlier section, which are based on real satellite and snow data. The authors should consider omitting it, and returning to these considerations in a future publication.

Response 5: We agree and deleted it.

6. Also, the authors should discuss the difference between snow depth and snow water equivalent (SWE). To my understanding, SWE is more relevant for hydrologic applications, and may be more directly measured by the microwave retrievals.

Response 6: We agree with the reviewer's opinion. Snow water equivalent (SWE), describing the amount of water stored in a snowpack, is a key variable for hydrological applications. Generally, a reasonable 'global' snow density (240 kg/m^3) is used to transfer snow depth to SWE (Takala et al., 2011).

In our study, we used the RF algorithm to retrieve snow depth rather than SWE because that station observations include only snow depth data.

Generally, snow density presents a variation in space and time. Thus, a relation to SWE through a fixed snow density is unreasonable. In the future, the temporospatial distribution of snow density in China will be mapped based on the reanalysis data from ERA5-land to improve SWE estimation. We are now assessing the ERA5 data using ground truth observations.

Takala, M., Luojus, K., Pulliainen, J., Lemmetyinen, J., Juha-Petri, K., Koskinen, J., and Bojkov, B., 2011. Estimating northern hemisphere snow water equivalent for climate research through assimilation of space-borne radiometer data and ground-based measurements. Remote Sensing of Environment. 115, 3517-3529.

7. On a related note, the authors note that snow measurements in high mountain areas are sparse, so that remote sensing based snow estimates cannot be validated. This could be partly overcome using a mass balance approach based on, for example, spring and summer streamflow measurements, which would give SWE (and hence, making assumptions about density, also snow depth) on a watershed scale (which in some cases might even be comparable with the satellite spatial resolution scale). See, e.g., Dahri et al.

(2018) "Adjustment of measurement errors to reconcile precipitation distribution in the high-altitude Indus basin" and related work.

Response 7: We appreciate your constructive suggestions. We are considering a snow depletion curve, e.g., Parallel Energy Balance Model, to improve the snow depth retrievals in high-altitude areas. We read the reference carefully and cited it in the revised manuscript. “Snow depth estimation in the mountains remains a challenge (Lettenmaier et al., 2015; Dozier et al., 2016; Dahri et al., 2018)” (Page 10, Line 25-26).

We would like to thank the four referees and the editor for dedicating their time to our manuscript and providing us with positive and constructive comments. We have studied the comments carefully and have made detailed corrections:

Given the extensive changes in the revised manuscript, here we make a summary of the main revisions:

- **Independent validation dataset:**

  One of major issues of the original study was that the validation data were not temporally and spatially independent from the training data. Thus, available stations were randomly divided into two roughly equally sized parts: training stations and validation stations. 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 reconstructed snow depth product.

- **Optimal input predictor variables for RF model:**

  The procedure described in the original manuscript was complicated due to so many predictor variables. Based on the correlations between the predictor variables and the variable importance metrics, 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. The scheme four was the most complicated. The predictor variables were satellites observations, latitude, longitude, elevation and land cover fraction. These four combinations of predictor variables, together with snow depth measurements, trained the four RF algorithms. We validated these four fitted RF algorithms to determine whether certain predictor variables are necessary and whether their inclusion affects the RF model.

- **Validation of the fitted RF algorithms:**

  The fitted RF algorithm was validated using temporally independent data in the original manuscript. To assess the feasibility of RF model in estimating snow depth, we conducted three tests to verify the fitted RF algorithms in the revised manuscript. The same training samples (same algorithms) were used for 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. 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 Test 3, a spatially independent dataset from validation stations during the period 2015-2018 was used to assess the accuracy of spatio-temporal prediction.

  According to the validation of the fitted RF algorithms, we found many redundant inputs due to highly correlated predictor variables. Thus, we used a straightforward fitted RF algorithm (trained with T_B and geolocation) to retrieve a consistent 32-year daily snow depth dataset from 1987 to 2018.

- **Validation of the reconstructed snow depth product:**

This product was evaluated against the independent station observations during the period 1987-2018. We also compared the performances of snow depth product in three snow cover areas over China.

- **Trends analysis of snow depth:**

  Three long-term (1987-2018) datasets, including ground truth observations, RF estimates and former snow depth product in China, were applied to analyze the trends of snow depth variation in China using the Mann-Kendall test and slope method.

- **Available long-term snow depth dataset:**

  The reconstructed dataset from 1987 to 2018 is now available and we will upload the data later.

- **Rewritten, simplified and better structured:**

  We revised the manuscript carefully and thoroughly to clarify the structure and content of the paper. We rewrote the results and discussion sections and split them. Additionally, a thorough revision of the manuscript was completed by a native speaker.

Snow Depth Estimation and Historical Data Reconstruction Over China Based on a Random Forest Machine Learning Approach

Jianwei Yang[1], Lingmei Jiang[1], Kari Luojus[2], Jinmei Pan[3], Juha Lemmetyinen[2], Matias Takala[2], Shengli Wu[4]

[1]State Key Laboratory of Remote Sensing Science, Jointly Sponsored by Beijing Normal University and the Institute of Remote Sensing and Digital Earth of Chinese Academy of Sciences, Beijing Engineering Research Center for Global Land Remote Sensing Products, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China

[2]Finnish Meteorological Institute, Helsinki Fi00101, Finland

[3]State Key Laboratory of Remote Sensing Science, Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing 100101, China

[4]National Satellite Meteorological Center, China Meteorological Administration, Beijing 100081, China

Corresponding Author: Lingmei Jiang (jiang@bnu.edu.cn)

Abstract. We studied whether the random forest (RF) machine learning (ML) model could be used to retrieve snow depth. Four combinations composed of critical predictor variables were used to train the RF model. Then, we utilized three validation datasets from out-of-bag (OOB) samples, a temporal subset and a spatiotemporal subset to verify the fitted RF algorithms. The results indicated the following: (1) the accuracy of RF model is greatly influenced by geographic location, elevation, and land cover fractions; (2) however, the redundant predictor variables (if highly correlated) slightly affect the RF model; (3) the fitted RF algorithms perform better in temporal scale than in spatial scale, with unbiased RMSEs of ~4.4 cm and ~7.3 cm, respectively. Finally, we 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). Although the RF product was superior to the WESTDC dataset, it still underestimated deep snow cover (> 20 cm), with biases of -10.4 cm, -8.9 cm and -34.1 cm in Northeast China (NE), northern Xinjiang (XJ) and the Qinghai-Tibetan Plateau (QTP), respectively. Additionally, the long-term snow depth datasets (station observations, RF estimates and WESTDC product) were analyzed in terms of temporal and spatial variations over China. On a temporal scale, the ground truth snow depth presented a significant increasing trend from 1987 to 2018, especially in NE. However, the RF and WESTDC products displayed no significant changing trends except in QTP. The WESTDC product presented a significant decreasing trend in QTP, with a correlation coefficient of -0.55, whereas there were no significant trends for ground truth observations and the RF product. For the spatial characteristics, similar trend patterns were observed for RF and WESTDC products over China. These characteristics presented significant decreasing trends in most areas and a significant increasing trend in central NE.

1 Introduction

Seasonal snow covers a considerable portion of the land surface in the Northern Hemisphere during winter and has a significant effect on the Earth's radiation balance and surface-atmosphere interaction due to its high albedo and low thermal conductivity (Fernandes et al., 2009; Derksen et al., 2012; Kevin et al., 2017; Dorji et al., 2018; Bormann et al., 2018). Snow depth is a crucial parameter for climate studies, hydrological applications and weather forecasts (Foster et al., 2011; Takala et al., 2017; Tedesco et al., 2016; Safavi et al., 2017). For these applications, long time series are needed to conduct meaningful statistics on trends and variability. Fortunately, passive microwave (PMW) signals can penetrate snow cover and provide snow depth estimates through volume scattering of snow particles in dry snow conditions. PMW remote sensing also has the advantage of sensing without depending on solar illumination and weather conditions (Chang et al., 1987; Foster et al., 2011). In addition, there exists a long historical record of spaceborne PMW data dating back to 1978, allowing us to study seasonal snow climatological changes (Takala et al., 2011; Santi et al., 2012). These advantages make snow depth estimation from satellite PMW remote sensing an attractive option.

Diverse methods have been proposed to retrieve snow depth from PMW observations. The most widely used inversion algorithms were based on empirical relationships between satellite brightness temperature (T_B) gradient and snow depth (Chang et al., 1987; Foster et al., 1997; Derksen et al., 2005; Che et al., 2008; Kelly et al., 2003; Kelly et al., 2009; Jiang et al., 2014). However, these algorithms are not always reliable in all regions due to the fixed empirical constants (Derksen et al., 2010; Davenport et al., 2012; Che et al., 2016; Yang et al., 2019). Subsequently, more advanced algorithms that use theoretical or semi-empirical radiative transfer models were developed (Jiang et al., 2007; Takala et al., 2011; Picard et al., 2012; Lemmetyinen et al., 2015; Metsämäki et al., 2015; Tedesco et al., 2016; Pan et al., 2017; Saberi et al., 2017); however, these complicated algorithms are computationally expensive and require complex ancillary data to provide accurate predictions. These factors restrict the applications of these algorithms on a global scale. Improving the performance of PMW retrieval algorithms through data assimilation has also been investigated (Durand et al., 2006; Tedesco et al., 2010; Che et al., 2014; Huang et al., 2017). The widely used and operational assimilation system combines synoptic weather station data with satellite PMW radiometer measurements through the snow forward model (Helsinki University of Technology snow emission model, HUT), and it provides long-term snow water equivalent data from 1979 to the present in the Northern Hemisphere (> 35º N) (Pulliainen et al., 1999; Pulliainen., 2006; Takala et al., 2011). However, the coverage of this product does not include the Qinghai-Tibetan Plateau (QTP), which is one of three stable snow cover areas in China.

Machine learning (ML) has attained outstanding results in the regression estimation of land surface parameters from remotely sensed observations at local and global scales over the past decade (Reichstein et al., 2019). The random forest (RF) is an ensemble method whereby multiple trees are grown from random subsets of predictors, producing a weighted ensemble of trees (Breiman, 2001). RF is also robust against overfitting in the presence of large datasets and increases predictive accuracies over single decision trees (Biau and Scornet, 2016; Tyralis et sl., 2019b). Over the last two decades, RF has been one of the most successful ML algorithms for practical applications due to its proven accuracy, stability, speed of processing and ease of use (Rodriguez-Galiano et al., 2012; Belgiu et al., 2016; Maxwell et al., 2018; Bair et al., 2018; Qu et al., 2019; Reichstein et al., 2019; Tyralis et sl., 2019a). Although the RF model can present good results in many research areas, studies on the spatio-temporal prediction of snow depth are few and the potential utility of RF in such studies is unknown.

The primary objectives of this study are to assess the feasibility of the RF model in estimating snow depth, to determine whether the inclusion of auxiliary information (geolocation, elevation and land cover fraction) contributes to the improvement of RF, and eventually to develop a time series (1987 to 2018) of snow depth data in China and analyze the trends in annual mean snow depth. To complete the feasibility study of the RF model, we designed four RF algorithms trained with different combinations of predictor variables and validated them using temporally and spatially independent reference data. To our knowledge, this type of assessment of RF algorithm performance has not been made to date over China. The data and methodology are described in Section 2. Section 3 presents the results regarding the feasibility study of the RF model, the validation of the snow depth product reconstructed with the RF algorithm and the trend analysis of snow depth. The results are discussed in Section 4, and conclusions are given in Section 5.

2 Data and Methodology

2.1 Data

(1) Satellite passive microwave measurements

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. The data are available from the National Snow and Ice Center (https://daacdata.apps.nsidc.org/pub/DATASETS). The SSM/I and SSMIS sensors are suitable for producing a long-term consistent snow depth dataset due to their similar configurations and intersensor calibrations (Armstrong et al., 1994). To avoid the influence of wet snow, only ascending (F08) and descending (F11, F13 and F17) overpass data were used (Table 1). In this study, the difference between 19.35 (36.5) GHz and 18.7 (37) GHz was ignored (hereafter referred as 19 GHz and 37 GHz, respectively).

(2) In situ measurements

The weather station daily data in China from 1987 to 2018 were provided by the National Meteorological Information Centre, China Meteorology Administration (CMA, http://data.cma.cn/en). The geographical locations of the meteorological stations and the three stable snow cover areas are shown in Fig. 1. The recorded variables include the site name, observation time, geolocation (latitude and longitude), altitude (m), near-surface soil temperature (measured at a 5-cm depth, °C), and snow depth (cm). The sites are not distributed homogeneously, and few are located in inaccessible regions with extreme climates and complex terrain conditions, e.g., the western part of QTP (Fig. 1).

Quality control was conducted prior to using the data for developing and validating the retrieval algorithm. The first step was to select the records where the near-surface soil temperature was lower than 0 °C. The second step was to remove the sites if the areal fraction of the open water exceeded 30% within a satellite pixel. Finally, the 683 stations were randomly divided into two roughly equal-sized parts (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 independent data spatially, can be applied to validate the fitted RF algorithm. 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.

(3) Land cover fraction

A 1-km land use/land cover (LULC) map derived from the 30-m Thematic Mapper (TM) imagery classification was provided by the Data Center for Resources and Environmental Sciences, Chinese Academy of Sciences (http://www.resdc.cn/). The map was recalculated as the areal percentages of each land cover type in the 25-km grid cells. In this study, the fractions of grassland, bareland, cropland, forest, and shrubland were calculated as predictor variables of the RF model. To avoid the influence of water bodies and construction, the record was used only if the total fraction was greater than 60%.

2.2 Methodology

2.2.1 Random forest

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).

2.2.2 Feasibility study of the RF model

(1) Selection of predictor variables

The possible predictor variables used include geographic location (longitude, latitude), elevation, land cover fractions (grassland, cropland, bareland, shrubland and forest) and multi-channel brightness temperatures. All available channels on the SSM/I and SSMIS are listed in Table 1. The 23 GHz channel is sensitive to water vapor and not surface scattering, which introduces uncertainty to the estimation process (Ji et al., 2017). The 85 (91) GHz channel is seriously influenced by the atmosphere (Kelly et al., 2009; Xue et al., 2017). Typically, the lower frequency (19 GHz) is used to provide a background T_B against which the higher frequency (37 GHz) scattering-sensitive channels are used to retrieve snow depth. The mixed-pixel problem is the dominant limitation on snow depth estimation accuracy (Derksen et al., 2005; Jiang et al., 2014; Roy et al., 2014; Cai et al., 2017; Li et al., 2017). The satellite pixel usually covers several land cover types due to a coarse footprint. Thus, the land cover fractions were included as possible predictor variables. Previous studies have shown that geographic location and elevation indeed contribute to improving ML model performance (Bair et al., 2018; Qu et al., 2019).

To determine a suitable selection rule for training samples, we selected four combinations of predictor variables from training stations (Fig. 1) during the period 2012-2014 to train the RF algorithms. Table 2 presents a detailed description of the four selection rules of training samples. The correlations between the predictor variables and the variable importance metrics are shown in Fig. 3. The T_B measurements at horizontal polarization (H-pol) are highly correlated (correlations higher than 0.9) with observations at vertical polarization (V-pol). Moreover, according to their ranking of the predictor variables, the channels of V-pol are more relevant to the independent variable (snow depth) than are the H-pol channels. Therefore, the RF1 algorithm was trained with only two channels' T_B measurements at V-pol. The ranking of variables' importance in Fig. 3 indicates that the geographic location is more important than elevation to snow depth. Thus, the geographic location and elevation were included in the predictor variables of RF2 and RF3, respectively. Fig. 3 also shows that the correlations between T_B and land cover fraction are relatively low. Thus, we will validate whether the inclusion of land cover fraction would increase the performance of the fitted RF4 algorithm.

(2) Training sample size

One of the advantages of the RF model is that it can effectively handle small sample sizes (Biau and Scornet et al., 2016). A test was conducted to demonstrate the insensitivity of the RF model to the training sample size. The input predictor variables include geographic location and T_B (Table 2, RF2). The flowchart of the test process is shown in Fig. 4. To ensure a sufficient number of samples, 80,000 records from 1987 to 2004 were used to test the required size of the training samples and a two-year stand-alone dataset from (2005-2006) was applied to assess the performance. During this process, the number of samples selected randomly was from 5000 to 80,000 (step, 5000). We consider three evaluating indicators (the unbiased root mean square error (RMSE), bias and correlation coefficient) to illustrate the sensitivity of the RF model to the training sample size.

(3) Validation datasets of the fitted RF algorithms

We conducted three tests to verify the fitted RF algorithms (Table 3). The same training samples (same algorithms) were used for the three tests but with different validation datasets. In Test1, the validation data were from OOB samples. This preliminary assessment generally offers a simple way to adjust the parameters of the RF model. However, the OOB errors should be used with caution because its samples are not independent at temporal and spatial scales. In Test2, we applied independent reference data during the period 2015-2018 to assess the accuracy of the temporal prediction of fitted algorithms. Although this dataset is composed of observations from training stations in Fig. 1, it is temporally independent of the training samples (2012-2014). Generally, the RF model cannot extrapolate outside the training range (Hengl et al., 2018). Thus, in Test3, a spatially independent dataset from validation stations during the period 2015-2018 was used to assess the accuracy of spatio-temporal prediction. The unbiased RMSE, bias and correlation coefficient are used for the assessment of the predictive performance of the fitted algorithms.

2.2.3 Validation of reconstructed snow depth product and trend analysis

The reconstructed long-term snow depth dataset was evaluated by the stand-alone ground truth measurements over the period 1987-2018 from the validation stations (Fig. 1). 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). Then, the spatiotemporal patterns of snow depth were analyzed in Northeast China (NE), northern Xinjiang (XJ), and the QTP. The slope method (regression) was employed to analyze the snow depth variation trend at the temporal scale (Huang et al., 2019). To show the spatial distribution of snow depth variation, the Mann-Kendall test (significance levels of α=0.05) was used to analyze the trends of changes in China (Mann., 1945; Kendall et al., 1975; Milan, 2013). To ensure the presence of dry snow cover, the reconstruction periods are the main snow winter season (January, February, March, November, and December).

3 Results

3.1 Sensitivity to training sample size

The sensitivity of the RF model toward the training sample size was evaluated to confirm the appropriate number of training samples. Fig. 5 displays the accuracy according to unbiased RMSE, bias, and correlation coefficient. These accuracy indexes show slight fluctuations when the number of training sample increases from 5000 to 80,000. Fig. 5a shows that the unbiased RMSE ranges from 5.1 cm to 5.5 cm with increasing training samples. Fig. 5c shows that the correlation coefficient is as high as 0.79 and becomes stable when the samples are up to 30,000. According to the sensitivity analysis, the number of training samples has less influence on the prediction accuracy of the RF model. This test is very helpful for us to determine the number of training samples because of the limited number of training samples over the period 2012-2014. We selected all available samples (28,602) from training stations (Fig. 1) during the period 2012-2014 to train the RF models in Table 2.

3.2 Validation of the fitted RF algorithms

The fitted RF algorithms were evaluated by three validation datasets as shown in Table 3. The color-density scatterplots of the measured snow depth versus the retrieved snow depth are presented in Fig. 6. For all fitted RF algorithms (RF1, RF2, RF3 and RF4), notable differences in accuracy were revealed through the validation of three datasets (Table 4). Generally, the validation with OOB samples presented higher overall accuracy than the other two datasets. This result, however, does not demonstrate that the fitted RF algorithm performs well in snow depth estimation. The assessments in Test2 (temporal subset) and Test3 (spatio-temporal subset) demonstrate that the temporal prediction of the RF model outperforms the spatio-temporal prediction, with unbiased RMSEs of 4.4-5.4 cm and 7.2-7.9 cm, respectively.

Comparing the validation results of RF1, RF2, RF3 and RF4, we find that the inclusion of auxiliary information indeed improved the performance of the fitted RF algorithms (Fig. 6). For Test1(OOB), the unbiased RMSE decreased from 6.4 cm to 3.9 cm with increasing predictor variables of auxiliary information, while the correlation coefficient increased from 0.72 to 0.90 (Table 4). For Test2(temporal subset), the unbiased RMSE decreased from 5.4 cm to 4.4 cm and the correlation coefficient increased from 0.77 to 0.85 (Table 4). There was a slight improvement in spatio-temporal prediction when including the auxiliary information, with the unbiased RMSE ranging from 7.9 cm to 7.3 cm (Table 4).

3.3 Validation of the reconstructed snow depth product

According to the results in Fig. 6 and Table 4, there are no notable differences in accuracy among the RF2, RF3, RF4 algorithms. In this study, we selected the RF2 algorithm to reconstruct a long-term snow depth dataset (1987 to 2018). We used the independent in situ measurements over the period 1987-2018 from validation stations (Fig. 1) to evaluate this product (hereafter, RF product). Fig. 7 shows the scatter diagrams of estimated vs. measured values for RF and WESTDC products. The overall accuracy of the RF product is higher than that of the WESTDC estimates, with unbiased RMSEs of 7.1 cm and 8.5 cm, respectively (Fig. 7a and 7b). The correlation coefficient is 0.65, which is larger than the WESTDC's coefficient of 0.49. Both products particularly underestimate snow depth when snowpack is thicker than 20 cm. The error bar shows that the WESTDC product tends to more seriously underestimate snow depth than do the RF estimates.

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. 8. The results show that the RF estimates outperform the WESTDC product with respect to unbiased RMSE and correlation coefficient from season to season. The bias also fluctuates from season to season, ranging from -8 cm to 3 cm (Fig. 8c). There is a slight overestimation during the period 1987-2000, whereas it presents a notable underestimation since 2006.

The assessment of snow depth product was performed in three snow cover areas of China. As shown in Fig. 9a, the RF data are superior to the WESTDC estimates, with the unbiased RMSEs of 8.3 cm, 6.8 cm and 8.8 cm in QTP, NE and northern XJ for the RF product, respectively. Fig. 9b shows a notable underestimation and overestimation for the WESTDC product in northern XJ and the QTP, respectively. For the RF product, the bias is close to zero and fluctuates across a relatively narrow range in the three snow cover areas.

Based on the results in Fig. 7, we selected 20 cm as a threshold to assess the performances in deep (> 20 cm) and shallow (≤20 cm) snow cover. The percentage of shallow snow conditions to total samples was approximately 90%. Table 5 displays the comparison between RF estimates and the WESTDC product in the three snow cover areas. The ‘Samples’ row in Table 5 shows the number of samples and the corresponding percentage in each region. Both products present notable underestimation for deep snow cover, with the biases of -34.1 cm and -33.8 cm in QTP for the RF and WESTDC products, respectively. The biases are -10.4 cm and -8.9 cm for the RF product in NE and northern XJ, respectively, whereas the same biases are -11.8 cm and -13.2 cm for the WESTDC data. Moreover, the correlation is very poor in deep snow cover, even negative (-0.18) in QTP for the WESTDC product. For shallow snow cover, the RF product is superior to the WESTDC estimates in QTP, with unbiased RMSEs of 3.4 cm (RF) and 5.6 cm (WESTDC). Furthermore, the WESTDC product presents overestimation in QTP, with a bias of 4.0 cm that is much higher than the RF’s bias of 0.6 cm. The unbiased RMSEs of the RF product are 5.4 cm and 6.1 cm in NE and northern XJ for shallow snow cover, respectively, lower than the WESTDC’s values of 6.5 cm and 7.4 cm. However, the RF product tends to overestimate snow depth relative to WESTDC estimates, with higher biases of 1.8 cm and 2.5 cm than WESTDC’s 0.5 cm and -0.4 cm in NE and northern XJ, respectively.

3.4 Spatial-temporal analysis of snow depth in three snow cover areas

The trend analysis of snow depth was conducted based on ground truth observations, the RF dataset and the WESTDC product during the period 1987-2018. The time series of yearly mean snow depth in different regions over China is shown in Fig. 10. The red, green and blue solid lines represent yearly mean snow depth in northern XJ, NE and QTP, respectively. The black solid line displays the overall mean snow depth in China. Fig. 10a shows that the ground truth snow depth in China presents a significant increasing trend from 1987 to 2018, with a correlation coefficient of 0.57. The trend in NE is highly consistent with the overall trend over China, with a correlation coefficient of 0.64 (Fig. 10a). Although there are increasing trends in northern

XJ and QTP, the correlation coefficients are lower than 0.40, not significant (Fig. 10a). Fig. 10b and 10c show the time series of yearly mean snow depth from the RF and WESTDC products, respectively. Neither of these values present significant trends. In the QTP, the WESTDC product presents a significant decreasing trend, with a correlation coefficient of -0.55 (Fig. 10c). Snow depth in northern XJ is the greatest among three snow cover areas, and snow cover in the QTP is very shallow, approximately 5 cm (Fig. 10a and 10b). With respect to magnitude and change trends, the ground truth observations and RF estimates in this study are consistent.

Fig. 11 shows the spatial patterns of snow depth variation based on the RF and WESTDC products. Only the area with continuous snow depth measurements from 1987 to 2018 is shown in Fig. 11. The two products show similar patterns in the most areas over China. There are notable trend differences between RF and WESTDC products in the northeast of QTP and western NE. The RF product presents an increasing trend in the northeast of QTP, whereas a significant decreasing trend is presented for the WESTDC product (Fig. 11a and 11b). In the western NE, there is a significant increasing for the RF product but no significant trend for WESTDC data.

Based on the comparison of trends in Fig. 11 and available station observations in Fig. 1, we selected two specific areas (black and green grids in Fig. 11) to test the changing trend. Fig. 12 shows the trends of snow depth based on the station observations (black solid line), RF estimates (red solid line) and WESTDC product (blue solid line). The ground truth snow depth presents a significant increasing trend in the specific area of NE, with a high correlation coefficient of 0.75 (Fig. 12a). The RF product shows a significant increasing trend, which is consistent with the ground truth data (Fig. 11a and Fig. 12a). Fig. 12b shows that WESTDC product displays a decreasing trend in the selected area of QTP, while station observations and RF estimates present no significant trends.

4 Discussion

4.1 Disadvantages of the RF model

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.

4.2 Influence of predictor variables on the RF model

Fig. 6 and Table 4 indicate that the inclusion of correlated predictor variables has a very slight influence in the predictive performance. Geographic location contributes to improving the RF model's temporal and spatio-temporal estimates, and the inclusion of both elevation and land cover fraction does not further improve the performance of the fitted models (Fig. 6). This is because elevation is highly correlated (correlations higher than 0.9) with geographic location (Fig. 3). Fig. 3 also indicates that the correlation between longitude or elevation and land cover type (e.g., grassland, cropland, forest and bareland) is significant. However, this correlation does not mean that the effects of elevation and land cover fraction on fitted RF model can be ignored. We tested the RF algorithms trained with T_B and elevation or land cover fraction data. The results (not shown here) indicate that these auxiliary data do improve the performance of the fitted algorithms. Strongly correlated variables have a very slight influence on the predictive performance of the RF model (Boulesteix et al. 2012). Therefore, in some cases, a few representative predictor variables should be selected.

4.3 Potential errors of the reconstrued snow depth

Fig. 7 indicates that the RF model does not fully solve the overestimation and underestimation problems. For deep snow (> 20 cm), the biases are up to -8.9 cm and -10.4 cm in NE and northern XJ, respectively. Deep snow conditions account for roughly 10% of all training samples (Fig. 2). The estimates for deep snow cover in the QTP exhibit a large bias of -34.1 cm. Fig. 6 also illustrates that the fitted RF algorithms have no predictive ability for extremely deep snow conditions, especially in QTP. We checked the training data and found that the extreme high snow depth data (> 60 cm) occurred in QTP. However, the number of such cases is very small. In addition, the station measurements are point values while the satellite grids have a spatial resolution of 25 km × 25 km. Thus, the representativeness of these data is questionable. Snow depth estimation in the mountains remains a challenge (Lettenmaier et al., 2015; Dozier et al., 2016; Dahri et al., 2018). Numerous studies have been conducted on the snow cover over the QTP and have indicated that the snow cover in the Himalayas is higher than elsewhere, ranging from 80% to 100% during the winter (Basang et al., 2017; Hao et al., 2018). Additionally, Dai et al. (2018) showed that deep snow (greater than 20 cm) was mainly distributed in the Himalayas, Pamir, and Southeastern Mountains. Thus, the RF product produced in this paper has poor performance in QTP for the deep snow cover.

Table 5 indicates that there is overestimation in NE and northern XJ for shallow snow cover, which may be due to the following reasons. First, the PMW signals are insensitive to thin snow cover, especially for fresh snow with low snow density and snow grain size. Second, the large diurnal temperature range tends to subject the snowpack to frequent freeze-thaw cycles and leads to rapid snow grain (~2 mm) and snow density (200-350 kg/m^3) growth and consequently a high T_B difference (Meløysund et al., 2007; Durand et al., 2008; Yang et al., 2015; Dai et al., 2017). Third, frozen soil reduces the accuracy of estimates. Both snow and frozen ground are volume-scattering materials, and they have similar microwave radiation characteristics, making them difficult to distinguish. In addition, a limiting factor in estimating snow depth for PMW remote sensing is the presence of liquid water. In this study, a snow cover detection method is used to filter out wet snow cover; however, there are still misclassification errors, especially at the end of the winter season (Grody and Basist., 1996; Liu et al., 2018). In such cases, satellite observations are mainly associated with the emissions from the wet surface of the snowpack. Therefore, in wet snow conditions, snow depth retrieval is not possible (Derksen et al., 2010; Tedesco et al., 2016).

5 Conclusions

The present study analyzed the application of the RF model to snow depth estimation at temporal and spatial scales. Temporally and spatially independent datasets were applied to verify the fitted RF algorithms. The results suggested that the accuracy of fitted RF algorithms was greatly influenced by auxiliary data, especially the geographic location. However, the inclusion of strongly correlated predictor variables (elevation and land cover fraction) did not further improve the RF estimates. Therefore, in some cases, a few representative predictor variables should be selected. Due to naive extrapolation outside the training range, the transferability of a fitted RF algorithm at the temporal scale was better that that in spatial terms, e.g., with unbiased RMSEs of 4.5 cm and 7.2 cm for the RF2 algorithm, respectively.

In this study, the fitted RF2 algorithm was used to retrieve a consistent 32-year daily snow depth dataset from 1987 to 2018. Then, an evaluation was carried out using independent reference data from the validation stations during the period 1987-2018. The overall unbiased RMSE and bias were 7.1 cm and -0.05 cm, respectively, outperforming the WESTDC product (8.4 cm and -1.20 cm). In QTP, the unbiased RMSE and bias of RF estimates for shallow (≤ 20 cm) snow cover were 3.4 cm and 0.59 cm, respectively, much lower than WESTDC's 5.6 cm and 4.02 cm. In NE and northern XJ, RF estimates were superior to the WESTDC product but still presented an underestimation for deep snow (> 20 cm), with biases of -10.4 cm and -8.9 cm, respectively.

Three long-term (1987-2018) datasets, including ground truth observations, RF estimates and WESTDC product, were applied to analyze the trends of snow depth variation in China. The results suggested that there existed different trends among the three datasets. The overall trend of snow depth in China presented a significant increasing based on the ground truth observations, with a correlation coefficient of 0.57. Moreover, the trend in NE was highly consistent with the overall trend in China, with a correlation coefficient of 0.64. Neither the WESTDC nor the RF product presented significant trends except in QTP. The WESTDC product showed a significant decreasing trend in QTP, with a correlation coefficient of -0.55, whereas there were no significant trends for ground truth observations and the RF product.

As discussed in Section 4, our reconstructed snow depth estimates are still challenged by several problems, e.g., underestimation for deep snow. Additional prior knowledge of snow cover, such as snow cover fraction, snow density, and snow grain size, is necessary to improve the RF model. Combining the snow forward model with the ML method will be the focus of future work. Furthermore, the mass balance approaches, e.g., the Parallel Energy Balance model, will be used to improve the snow depth retrievals in high-altitude areas. In addition, although our results indicate that the RF method is a promising potential tool for snow depth estimation, there are a few pitfalls such as the risk of naive extrapolation and poor transferability in spatial terms limiting its application in spatio-temporal dynamics. It is in addressing these shortcomings that the techniques of deep learning promise breakthroughs. We are attempting to operate the Deep Neural Networks (DNN) model to overcome the limitations of traditional ML approaches.

Author contributions. L. Jiang conceived and designed the study; J. Yang produced the first draft of the manuscript, which was subsequently edited by J. Lemmetyinen, K. Luojus, L. Jiang and J. Pan; and K. Luojus, M. Takala, S. Wu, J. Pan and J. Yang contributed to the analytical tools and methods.

Competing interests. The authors declare that they have no conflicts of interest.

Acknowledgments. This study was supported by the Science and Technology Basic Resources Investigation Program of China (2017FY100502) and the National Natural Science Foundation of China (41671334). The authors would like to thank the China Meteorological Administration, National Geomatics Center of China, National Snow and Ice Data Center and NASA's Earth Observing System Data and Information System for providing the meteorological station measurements, land cover products and satellite datasets.

Data availability. Satellite passive microwave measurements are available for download from https://nsidc.org/. The in situ measurements provided by the China Meteorology Administration (CMA) and Chinese Snow Survey (CSS) project are not available to the public due to legal constraints. The land use/land cover (LULC) map was provided by the Data Center for Resources and Environmental Sciences, Chinese Academy of Sciences (http://www.resdc.cn/). The daily snow depth product was obtained from the Environmental and Ecological Science Data Center for West China (http://westdc.westgis.ac.cn).

References

Armstrong, R., Knowles, K., Brodzik, M., and Hardman, M.: DMSP SSM/I-SSMIS Pathfinder Daily EASE-Grid Brightness Temperatures, Version 2. Boulder, Colorado USA. NASA National Snow and Ice Data Center Distributed Active Archive Center, 10.5067/3EX2U1DV3434, 1994.

Bair, E. H., Abreu Calfa, A., Rittger, K., and Dozier, J.: Using machine learning for real-time estimates of snow water equivalent in the watersheds of Afghanistan, The Cryosphere, 12, 1579-1594, 10.5194/tc-12-1579-2018, 2018.

Basang, D., Barthel, K., Olseth, JA.: Satellite and Ground Observations of Snow Cover in Tibet during 2001–2015, Remote Sensing, 9,1201,10.3390/rs9111201, 2017.

Belgiu, M., and Lucian, D.: Random forest in remote sensing: A review of applications and future directions, ISPRS Journal of Photogrammetry and Remote Sensing, 114, 24-31. 10.1016/j.isprsjprs.2016.01.011, 2016.

Breiman, L., Cutler, A., Liaw, A., Wiener, M.: randomForest: Breiman and Cutler's Random Forests for Classification and Regression, R package version 4.6-14, 2018. https://CRAN.R-project.org/package=randomForest.

Bormann, K.J., Brown, R.D., Derksen, C., Painter, T.H.: Estimating snow-cover trends from space, Nat. Clim. Chang, 8, 924–928, 2018.

Biau, G.Ã.Š. and Scornet, E.: A random forest guided tour, TEST, 25, 197–227, 10.1007%2Fs11749-016-0481-7, 2016.

Breiman, L. Random forests. Mach. Learn. 2001, 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Chang, A., Foster J., Hall D.: Nimbus-7 derived global snow cover parameters, Annals of Glaciology, 9, 39-44, 10.1017/S0260305500000355, 1987.

Che, T., Dai, L., Zheng, X., Li, X., Zhao, K.: Estimation of snow depth from passive microwave brightness temperature data in forest regions of northeast China, Remote Sensing of Environment, 183, 334–349, 10.1016/j.rse.2016.06.005, 2016.

Che, T., Li, X., Jin, R., Armstrong, R., and Zhang, T.: Snow depth derived from passive microwave remote-sensing data in China, Annals of Glaciology, 49,145-154,10.3189/172756408787814690, 2008.

Che, T., Li, X., Jin, R., and Huang, C.: Assimilating passive microwave remote sensing data into a land surface model to improve the estimation of snow depth, Remote Sensing of Environment, 143, 54-63,10.1016/j.rse.2013.12.009, 2014.

Cai, S., Li, D., Durand, M., and Margulis, S.: Examination of the impacts of vegetation on the correlation between snow water equivalent and passive microwave brightness temperature, Remote Sensing of Environment, 193, 244–256, 10.1016/j.rse.2017.03.006, 2017.

Canovas-Garcia, F., Alonso-Sarria, F., Gomariz-Castillo, F., and Onate-Valdivieso, F.: Modification of the random forest algorithm to avoid statistical dependence problems when classifying remote sensing imagery, Comput. Geosci, 103, 1–11, 10.1016/j.cageo.2017.02.012, 2017.

Dahri, Z., Moors, E., Ludwig, F., Ahmad, S., Khan, A., Ali, I., Kabat, P.: Adjustment of measurement errors to reconcile precipitation distribution in the high-altitude Indus basin, Int J Climatol, 38, 1–19, 10.1002/joc.5539, 2018.

Dai, L., Che, T., Ding, Y., and Hao, X.: Evaluation of snow cover and snow depth on the Qinghai–Tibetan Plateau derived from passive microwave remote sensing, The Cryosphere, 11, 1933–1948, 10.5194/tc-11-1933-2017, 2017.

Dai, L., Che, T., Xie, H., and Wu, X.: Estimation of Snow Depth over the Qinghai-Tibetan Plateau Based on AMSR-E and MODIS Data, Remote Sensing, 10, 1989, 10.3390/rs10121989, 2018.

Davenport, I., Sandells, M., and Gurney, R.: The effects of variation in snow properties on passive microwave snow mass estimation, Remote Sensing of Environment, 118, 168–175, 10.1016/j.rse.2011.11.014, 2012.

Derksen, C., Walker, A., and Goodison, B.: Evaluation of passive microwave snow water equivalent retrievals across the boreal forest/tundra transition of western Canada, Remote Sensing of Environment, 96, 315-327, 10.1016/j.rse.2005.02.014, 2005.

Derksen, C., Toose, P., Rees, A., Wang, L., English, M., Walker, A., and Sturm, M.: Development of a tundra-specific snow water equivalent retrieval algorithm for satellite passive microwave data, Remote Sensing of Environment, 114, 1699–1709, 10.1016/j.rse.2010.02.019, 2010.

Derksen. C., and Brown, R.: Spring snow cover extent reductions in the 2008–2012 period exceeding climate model projections, Geophysical Research Letters, 39, 1-6, 10.1029/2012GL053387, 2012.

Dozier, J., Bair, E. H., and Davis, R. E.: Estimating the spatial distribution of snow water equivalent in the world's mountains, WIREs Water, 3, 461-474, doi 10.1002/wat2.1140, 2016.

Dorji, T., Hopping, K., Wang, S., Piao, S., Tarchen, T., and Klein, J.: Grazing and spring snow counteract the effects of warming on an alpine plant community in Tibet through effects on the dominant species, Agric. For. Meteorol, 263, 188–197, 10.1016/j.agrformet.2018.08.017, 2018.

Durand, M., and Margulis, S.: Feasibility test of multifrequency radiometric data assimulation to estimate snow water equivalent, Journal of Hydrometeorology, 7, 443-457, 10.1175/jhm502.1, 2006.

Durand, M., Kim, E., and Margulis, S.: Quantifying uncertainty in modeling snow microwave radiance for a mountain snowpack at the point-scale, including stratigraphic effects, IEEE Trans. Geosci. Remote Sens, 46, 1753–1767, 10.1109/tgrs.2008.916221, 2008.

Fernandes, R., Zhao, H., Wang, X., Key, J., Qu, X., and Hall, A.: Controls on Northern Hemisphere snow albedo feedback quantified using satellite Earth observations, Geophys. Res. Lett, 36, 1–6, 10.1029/2009gl040057, 2009.

Foster, J., Chang, A., Hall D.: Comparison of Snow Mass Estimation From a Prototype Passive Microwave Snow Algorithm, a Revised Algorithm and Snow Depth Climotology, Remote Sensing of Environment, 62, 132–142, 10.1016/S0034-4257(97)00085-0, 1997.

Foster, J., Hall, D., Eylander, J., Riggs, G., Nghiem, S., Tedesco, M., Kim, E., Montesano, P., Kelly, R., Casey, K., and Choudhury, B.: A blended global snow product using visible, passive microwave and scatterometer satellite data, International Journal of Remote Sensing, 32, 41 1371-1395, 10.1080/01431160903548013, 2011.

Grody, N., Basist, A.: Global identification of snow cover using SSM/I measurements, IEEE Trans. Geosci. Remote Sens, 34, 237–249, 10.1109/36.481908, 1996.

Hao, S., Jiang, L., Shi, J., Wang, G., Liu, X.: Assessment of MODIS-Based Fractional Snow Cover Products Over the Tibetan Plateau, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 99, 1-16, 10.1109/JSTARS.2018.2879666, 2018.

Hengl, T. et al.: SoilGrids250m: global gridded soil information based on machine learning. PLoS ONE 12, e0169748, 2017.

Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B.M., Gräler, B.: Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables, PeerJ, 10.7717/peerj.5518, 2018.

Huang, C., Newman, A., Clark M., Andrew, W., and Zheng, X.:Evaluation of snow data assimilation using the Ensemble Kalman Filter for seasonal streamflow prediction in the Western United States, Hydrology and Earth System Sciences, 21, 635-650, 10.5194/hess-21-635-2017, 2017.

Ji, D.B., Shi, J.C., Xiong, C., Wang, T.X., Zhang, Y.H.: A total precipitable water retrieval method over land using the combination of passive microwave and optical remote sensing, Remote Sensing of Environment, 191, 313-327, 2017.

Jiang, L., Shi, J., Tjuatja, S., Dozier, J., Chen, K., and Zhang, L.: A parameterized multiple-scattering model for microwave emission from dry snow, Remote Sensing of Environment, 111, 357-366, 10.1016/j.rse.2007.02.034, 2007.

Jiang, L., Wang, P., Zhang, L., Yang, H., Yang, J.: Improvement of snow depth retrieval for FY3B-MWRI in China, Science China: Earth Sciences, 44,531-47, 10.1007/s11430-013-4798-8,2014.

Kelly, R., Chang, A., Leung, T., and Foster, L.: A prototype AMSR-E global snow area and snow depth algorithm, IEEE Transactions on Geoscience and Remote Sensing, 41, 230 - 242, 10.1109/TGRS.2003.809118, 2003.

Kelly, R.: The AMSR-E Snow Depth Algorithm: Description and Initial Results, Journal of The Remote Sensing Society of Japan, 29, 307-317, 10.11440/rssj.29.307, 2009.

Kendall, M. G.: Rank Correlation Methods, Griffin, London, 1975.

Kevin, J., Kotlarski, S., Scherrer, S., and Schär, C.: The Alpine snow-albedo feedback in regional climate models, Climate Dynamics, 48, 1109–1124, 10.1007/s00382-016-3130-7, 2017.

Kühnlein, M., Appelhans, T., Thies, B. & Nauss, T.: Improving the accuracy of rainfall rates from optical satellite sensors with machine learning—a random forests-based approach applied to MSG SEVIRI, Remote Sens. Environ, 141,129–143, 2014.

Lemmetyinen, J., Derksen, C., Toose, P., Proksch, M., Pulliainen, J., Kontu, A., Rautiainen, K., and Seppänen, J.: Hallikainen, M. Simulating seasonally and spatially varying snow cover brightness temperature using HUT snow emission model and retrieval of a microwave effective grain size, Remote Sensing of Environment, 156, 71–95, 10.1016/j.rse.2014.09.016, 2015.

Lettenmaier, D., Alsdorf, D., Dozier, J., Huffman, G., Pan, M., and Wood, E.: Inroads of remote sensing into hydrologic science during the WRR era, Water Resour. Res, 51, 7309-7342, 10.1002/2015WR017616, 2015.

Li, Q., Kelly, R.: Correcting Satellite Passive Microwave Brightness Temperatures in Forested Landscapes Using Satellite Visible Reflectance Estimates of Forest Transmissivity, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 10, 3874-3883, 10.1109/JSTARS.2017.2707545, 2017.

Liaw, A., and Wiener, M.: Classification and regression by randomForest, R News, 2, 18–22, 2002.

Liu, X., Jiang, L., Wu, S., Hao, S., Wang, G., and Yang, J.: Assessment of Methods for Passive Microwave Snow Cover Mapping Using FY-3C/MWRI Data in China, Remote Sensing, 10, 524-539, 10.3390/rs10040524, 2018.

Liu, X., Jiang, L., Wang, G., Hao, S., and Chen, Z.: Using a Linear Unmixing Method to Improve Passive Microwave Snow Depth Retrievals, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens, 11, 4414–4429, 10.1109/PIERS.2016.7735542, 2018.

Maxwell, A., Warner, T., and Fang, F.: Implementation of machine-learning classification in remote sensing: An applied review, Int. J. Remote Sens, 39, 2784–2817, 2018.

Mann, H. B.: Nonparametric tests against trend, Econometrica 13, 245–259, 1945.

Milan, G., and Slavisa, T.: Analysis of changes in meteorological variables using Mann-Kendall and Sen's slope estimator statistical tests in Serbia, Global Planet Change, 100, 172-182, 10.1016/j.gloplacha.2012.10.014, 2013.

Meløysund, V., Bernt, L., Karl, V., and Kim R.: Predicting snow density using meteorological data, Meteorological Applications, 14, 413–423, 10.1002/met.40, 2007.

Metsämäki, S., Pulliainen, J., Salminen, M., Luojus, K., Wiesmann, A., Solberg, R., Böttcher, K., Hiltunen, M., and Ripper, E.: Introduction to GlobSnow Snow Extent products with considerations for accuracy assessment, Remote Sensing of Environment, 156, 96–108, 10.1016/j.rse.2014.09.018, 2015.

Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., Schaepman, M., Papritz, A.: Evaluation of digital soil mapping approaches with large sets of environmental covariates, Soil, 4, 1, 10.5194/soil-4-1-2018, 2018.

Orsolini, Y., Wegmann, M., Dutra, E., Liu, B., Balsamo, G., Yang, K., de Rosnay, P., Zhu, C., Wang, W., Senan, R., and Arduini, G.: Evaluation of snow depth and snow cover over the Tibetan Plateau in global reanalyses using in situ and satellite remote sensing observations, The Cryosphere, 13, 2221–2239, 10.5194/tc-13-2221-2019, 2019.

Pan, J., Durand, M., Vander Jaqt, B., and Liu, D.: Application of a Markov Chain Monte Carlo algorithm for snow water equivalent retrieval from passive microwave measurements, Remote Sensing of Environment, 192, 150-165, 10.1016/j.rse.2017.02.006, 2017.

Picard, G.: Simulation of the microwave emission of multi-layered snowpacks using the dense media radiative transfer theory: The DMRT-ML model, Geosci. Model Develop. Discuss., 6, 3647–3694, 2012.

Prasad, A., Iverson, L., and Liaw, A.: Newer classification and regression tree techniques: bagging and random forests for ecological prediction, Ecosystems, 9, 181-199, 10.1007/s10021-005-0054-1, 2006.

Probst, P., and Boulesteix, A.: To tune or not to tune the number of trees in random forest, J. Mach. Learn. Res, 18, 1–18, 2018.

Pulliainen, J., Grandell, J., and Hallikainen, M.: HUT snow emission model and its applicability to snow water equivalent retrieval, IEEE Trans. Geosci. Remote Sens, 37, 1378–1390,10.1109/36.763302, 1999.

Pulliainen, J.: Mapping of snow water equivalent and snow depth in boreal and sub-arctic zones by assimilating space-borne microwave radiometer data and ground-based observations, Remote Sens. Environ, 101, 257–269, 10.1016/j.rse.2006.01.002, 2006.

Qu, Y., Zhu, Z., Chai, L., Liu, S., Montzka, C., Liu, J., Yang, X., Lu, Z., Jin, R., Li, X., Guo, Z., and Zheng, J.: Rebuilding a Microwave Soil Moisture Product Using Random Forest Adopting AMSR-E/AMSR2 Brightness Temperature and SMAP over the Qinghai–Tibet Plateau, China, Remote Sensing, 11, 683, 10.3390/rs11060683, 2019.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat.: Deep learning and process understanding for data-driven Earth system science, Nature 566, 195–204, 2019.

Rodriguez-Galiano, V., Ghimire, B., Rogan, J., Chica-Olmo, M., and Rigol-Sanchez, J.: An assessment of the effectiveness of a random forest classifier for land-cover classification, ISPRS J. Photogramm. Remote Sens, 67, 93–104, 10.1016/j.isprsjprs.2011.11.002, 2012.

Roy, A., Royer, A., and Hall R.: Relationship Between Forest Microwave Transmissivity and Structural Parameters for the Canadian Boreal Forest, IEEE Geoscience and Remote Sensing Letters, 11, 1802-1806,10.1109/LGRS.2014.2309941, 2014.

Saberi, N., Kelly, R., Toose, P., Roy, A., and Derksen, C.: Modeling the observed microwave emission from shallow multi-layer tundra snow using DMRT-ML, Remote Sensing, 9, 1327, 10.3390/rs9121327, 2017.

Safavi, H., Sajjadi, S., and Raghibi, V.: Assessment of climate change impacts on climate variables using probabilistic ensemble modeling and trend analysis, Theoretical and Applied Climatology, 130, 635–653, 10.1007/s00704-016-1898-3, 2017.

Santi, E., Pettinato, S., Paloscia, S., Pampaloni, P., MacElloni, G., and Brogioni, M.: An algorithm for generating soil moisture and snow depth maps from microwave spaceborne radiometers: HydroAlgo, Hydrology and Earth System Sciences, 16, 3659–3676, 10.5194/hess-16-3659-2012, 2012.

Sturm, M., Holmgren, J., Liston, G.E.: A seasonal snow cover classification system for local to global applications, J. Clim, 8, 1261–1283, 1995.

Sturm, M., and Wagner, A.M.: Using repeated patterns in snow distribution modeling: An arctic example, Water Resour. Res, 46, 65–74, 2010.

Takala, M., Luojus, K., Pulliainen, J., Lemmetyinen, J., Juha-Petri, K., Koskinen, J., and Bojkov, B.: Estimating northern hemisphere snow water equivalent for climate research through assimilation of space-borne radiometer data and ground-based measurements, Remote Sensing of Environment, 115, 3517-3529, 10.1016/j.rse.2011.08.014, 2011.

Takala, M., Ikonen, J., Luojus, K., Lemmetyinen, J., Metsämäki, S., Cohen, J., Arslan, A., and Pulliainen J.: New Snow Water Equivalent Processing System With Improved Resolution Over Europe and its Applications in Hydrology, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 10, 428-436, 10.1109/JSTARS.2016.2586179, 2017.

Tedesco, M., and Narvekar, P.: Assessment of the NASA AMSR-E SWE product, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 3, 141-159, 10.1109/jstars.2010.2040462, 2010.

Tedesco, M., and Jeyaratnam, J.: A new operational snow retrieval algorithm applied to historical AMSR-E brightness temperatures, Remote Sensing, 8, 1037, 10.3390/rs8121037, 2016.

Tyralis, H., Papacharalampous, G., and Langousis, A.: A Brief Review of Random Forests for Water Scientists and Practitioners and Their Recent History in Water Resources, Water, 11, 910, 2019a.

Tyralis, H., Papacharalampous, G., and Tantanee, S.: 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, 10.1016/j.jhydrol.2019.04.070, 2019b.

Vaysse, K., and Lagacherie, P.: Evaluating digital soil Mapping approaches for mapping GlobalSoilMap soil properties from legacy data in Languedoc-Roussillon (France), Geoderma Regional, 4, 20-30, 10.1016/j.geodrs.2014.11.003, 2015.

Xue, Y., and Forman, B.A.: Atmospheric and Forest Decoupling of Passive Microwave Brightness Temperature Observations Over Snow-Covered Terrain in North America, IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing, 10, 3172–3189, 2017.

Yang, J., Jiang, L., Ménard, C., Luojus, K., Lemmetyinen, J., and Pulliainen, J.: Evaluation of snow products over the Tibetan Plateau, Hydrol. Processes, 29, 3247–3260, 10.1002/hyp.10427, 2015.

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.

Zhong, X., Zhang, T., Kang, S., Wang, K., Zheng, L., Hu, Y., and Wang, H.: Spatiotemporal variability of snow depth across the Eurasian continent from 1966 to 2012, The Cryosphere, 12, 227-245, 10.5194/tc-12-227-2018, 2018.

Ziegler, A., König, I.R.: Mining data with random forests: Current options for real-world applications, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 4, 55–63, 10.1002/widm.1114, 2014.

List of Tables and Figures

Table 1. 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

Table 2. 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, shrubland, 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		

Table 3. Summary of three tests of the fitted RF algorithms in Table 2.

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

Table 4. Accuracy of four snow-depth retrieval models with unbiased RMSE, bias and correlation coefficient.

Name	Test1 (OOB)			Test2 (temporal subset)			Test3 (spatio-temporal subset)		
	unRMSE	bias	corr.coe	unRMSE	bias	corr.coe	unRMSE	bias	corr.coe

RF1	6.4	-0.01	0.72	5.4	0.12	0.77	7.9	-0.76	0.57
RF2	4.1	0.07	0.90	4.5	0.27	0.85	7.2	-0.97	0.66
RF3	3.9	0.08	0.90	4.5	0.24	0.85	7.3	-0.83	0.66
RF4	3.9	0.03	0.91	4.4	0.21	0.85	7.3	-0.40	0.65

Table 5. Comparison between RF estimates and WESTDC product in three stable snow cover areas for deep (> 20 cm) and shallow (≤ 20 cm) snow cover.

RF product						
Regions	QTP		NE		northern XJ	
SnowDepth (cm)	<= 20	> 20	<= 20	> 20	<= 20	> 20
corr.coe	0.30	0.06	0.49	0.17	0.48	0.31
bias (cm)	0.59	-34.12	1.79	-10.38	2.52	-8.85
unRMSE (cm)	3.43	20.70	5.36	7.00	6.12	9.62
Samples	15503 (96.4%)	583 (3.6%)	151939 (87.3%)	22168 (12.7%)	32468 (69.8%)	14051 (30.2%)
WESTDC product						
Regions	QTP		NE		northern XJ	
SnowDepth (cm)	<= 20	> 20	<= 20	> 20	<= 20	> 20
corr.coe	0.16	-0.18	0.37	0.03	0.34	0.16
bias (cm)	4.02	-33.78	0.47	-11.75	-0.39	-13.22
unRMSE (cm)	5.60	21.62	6.47	9.10	7.35	11.30
Samples	15503 (96.4%)	583 (3.6%)	151939 (87.3%)	22168 (12.7%)	32468 (69.8%)	14051 (30.2%)

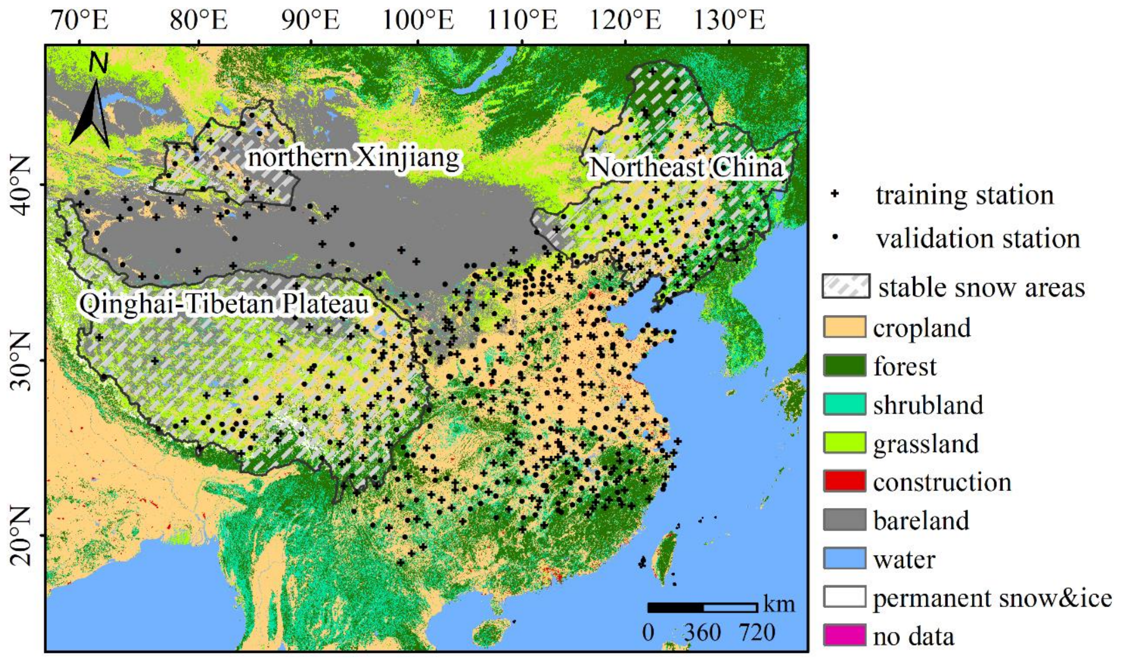

Figure 1. Spatial distribution of the weather stations and land cover types in the study area. There are three stable snow cover areas in China: Northeast China (NE), northern Xinjiang (XJ) and the Qinghai-Tibetan Plateau (QTP).

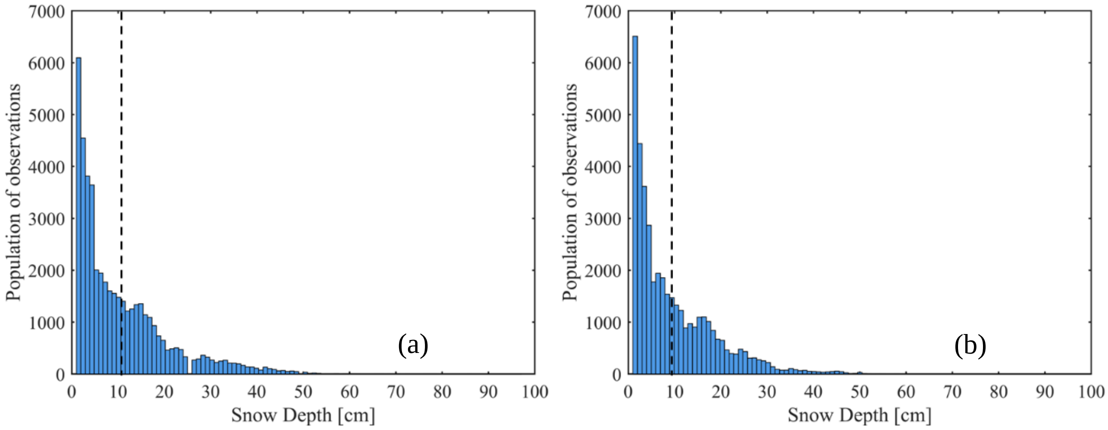

Figure 2. Histograms of snow depth observations from (a) training and (b) validation stations. The average values (black dashed lines) are equal to 10.5 cm and 9.8 cm, respectively.

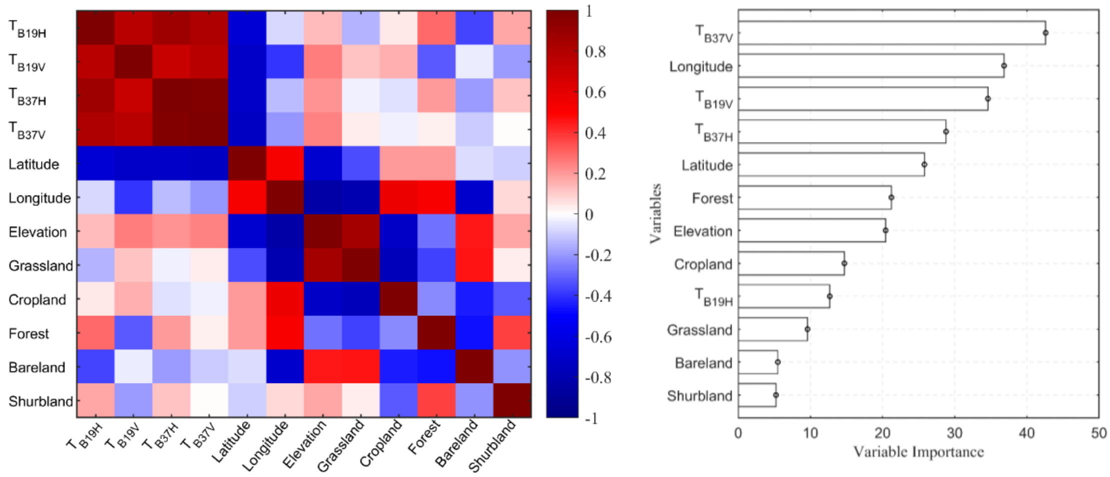

Figure 3. Correlations between the predictor variables (left) and the ranking of variable importance (right). The importance of variables, referred to as Mean Decrease Accuracy (MDA) in the RF model, is obtained by averaging the difference in out-of-bag error estimation before and after the permutation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

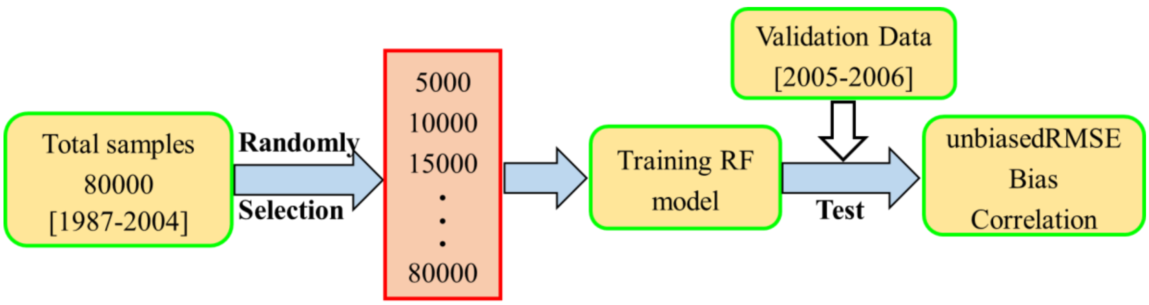

Figure 4. The test process flowchart for the sensitivity of the RF model to the training sample size.

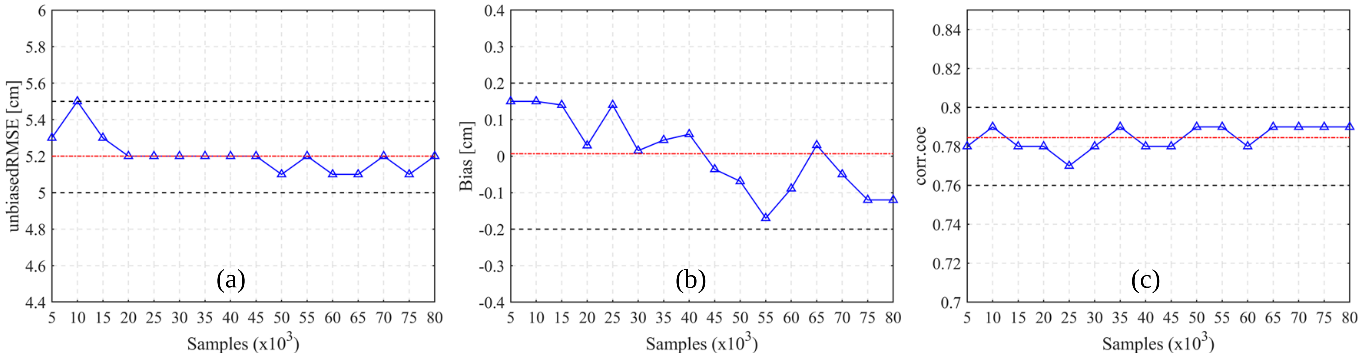

Figure 5. Trends of (a) unbiased RMSE, (b) bias and (c) correlation coefficient with increasing training sample size.

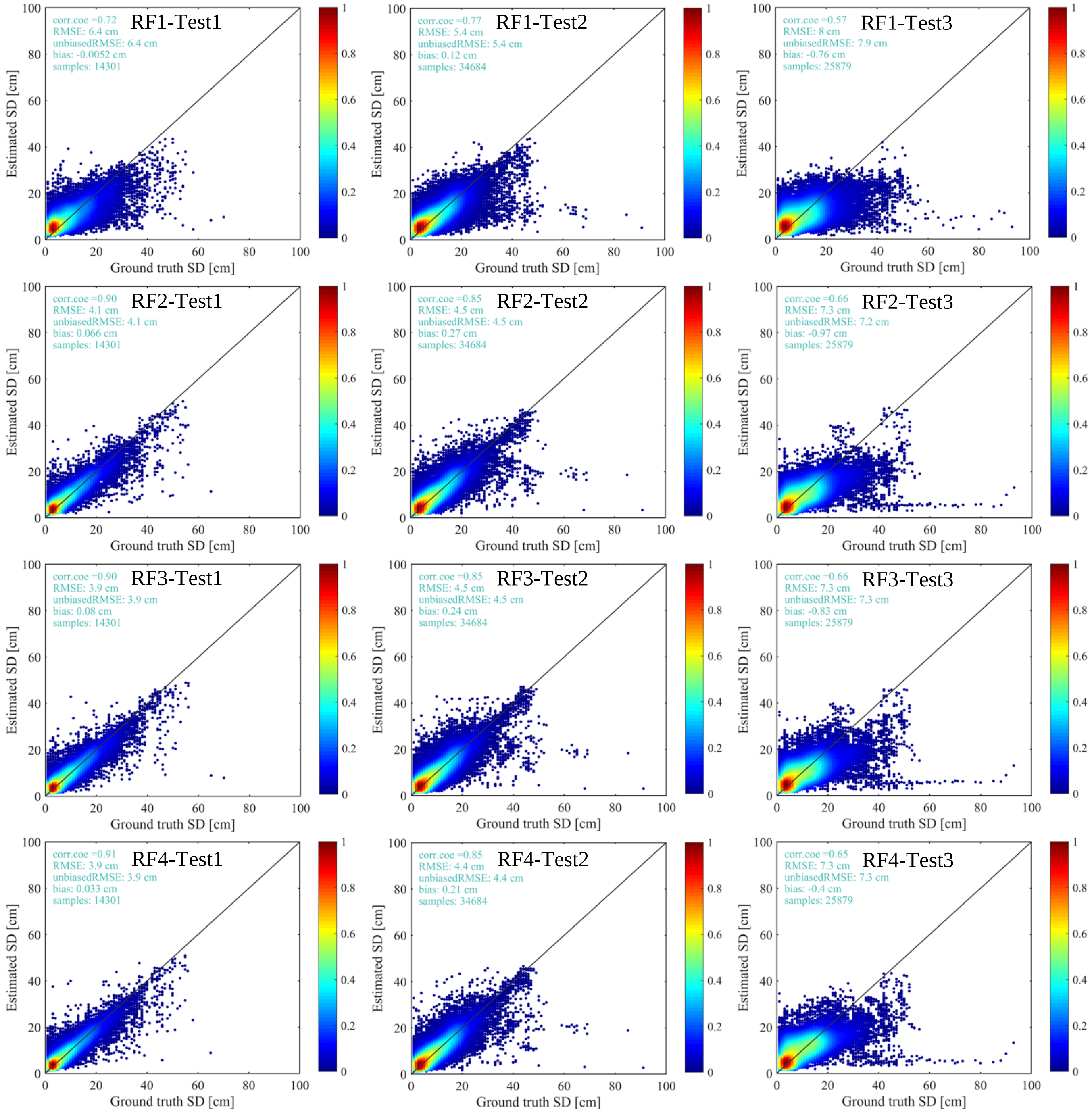

Figure 6. The color-density scatterplots of the estimated snow depth with four fitted RF algorithms and the ground truth snow depth. The four trained RF algorithms (RF1, RF2, RF3, RF4) were evaluated with three validation datasets (Test1, Test2, Test3).

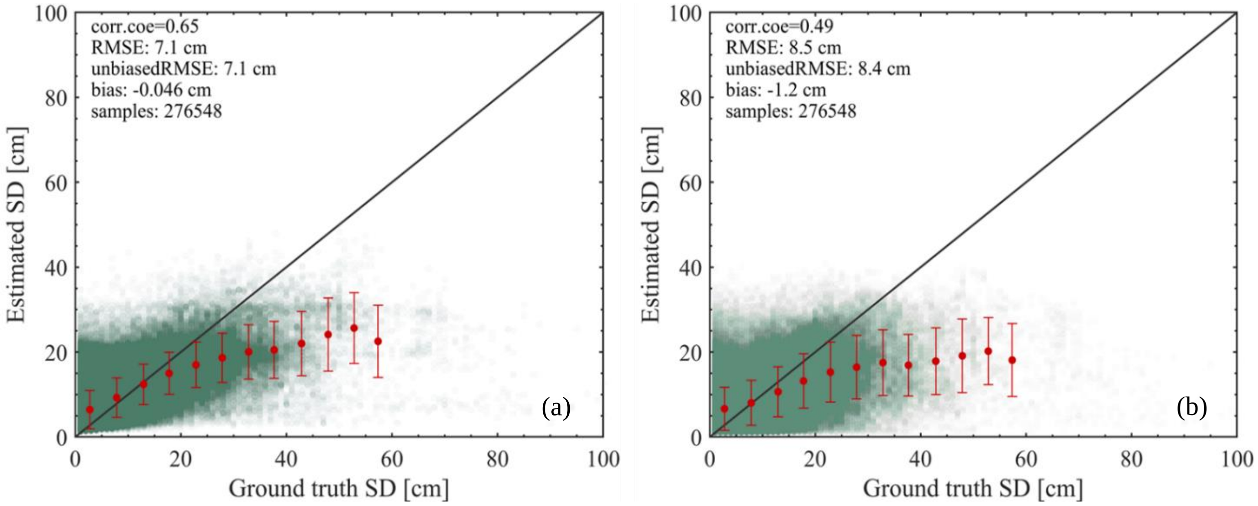

Figure 7. Scatterplots of the estimated snow depth and the ground truth observation for (a) RF and (b) WESTDC products.

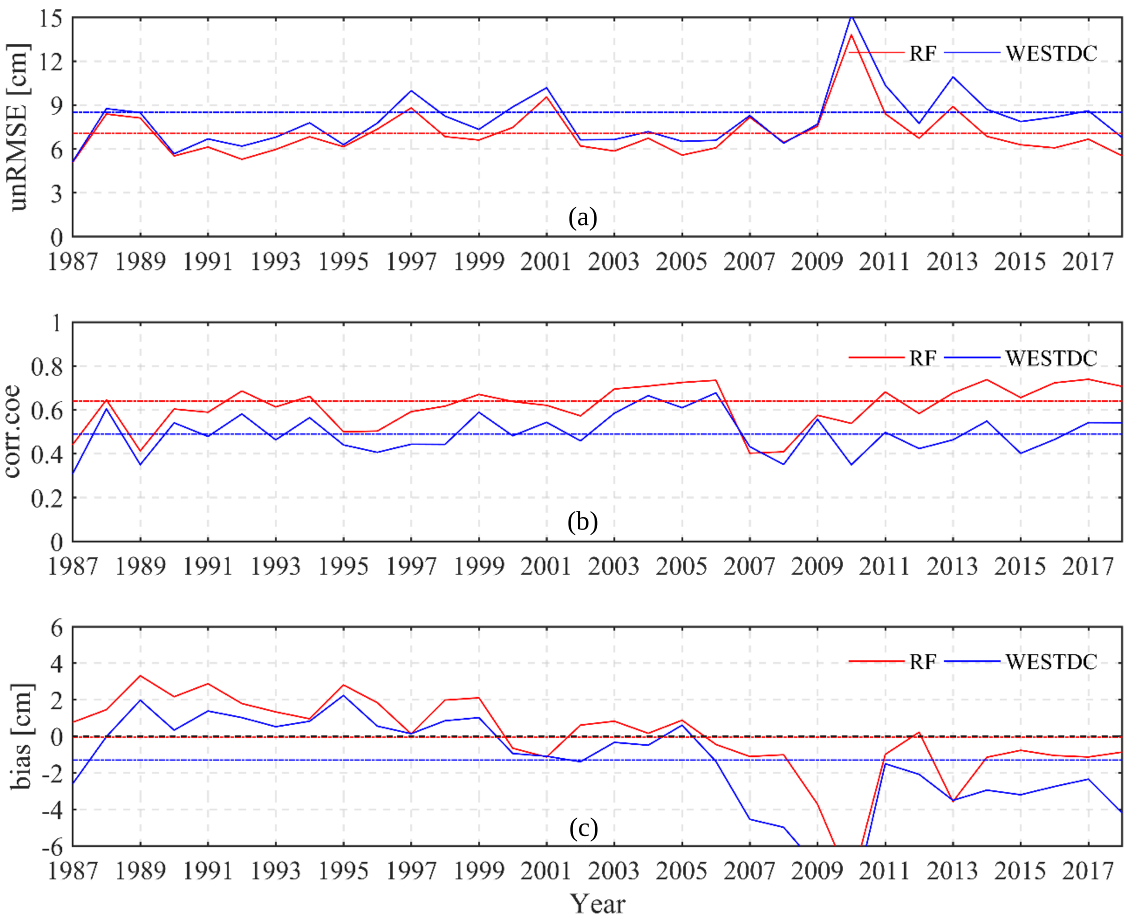

Figure 8. Time series of (a) unbiased RMSE (unRMSE), (b) correlation coefficient (corr.coe) and (c) bias for RF and WESTDC products. The colorful dashed lines represent mean values of assessment indexes.

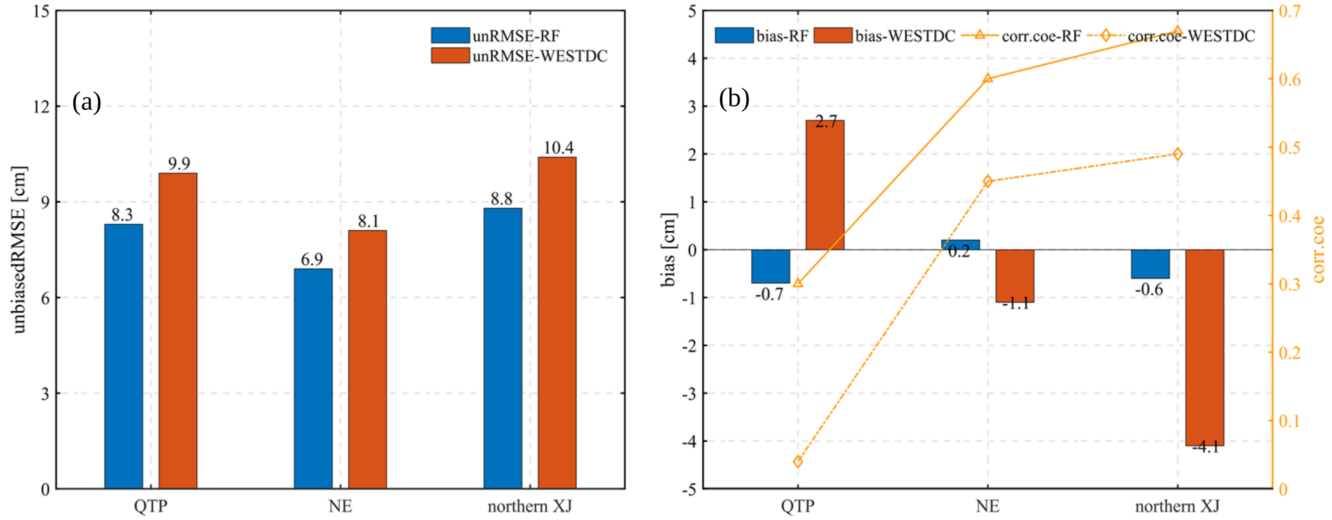

Figure 9. The validation of RF and WESTDC snow depth products in three stable snow cover areas over China with respect to (a) the unbiased RMSE, (b) bias and correlation coefficient.

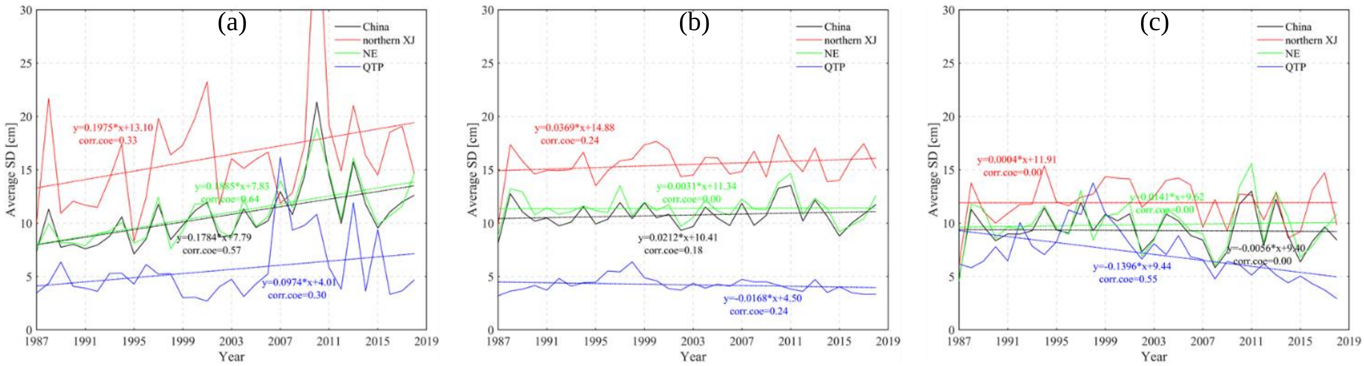

Figure 10. Trend analysis of snow depth based on (a) station observations, (b) RF estimates, and (d) WESTDC product in three stable snow cover areas of China. The correlation is statistically significant at the 0.05 level.

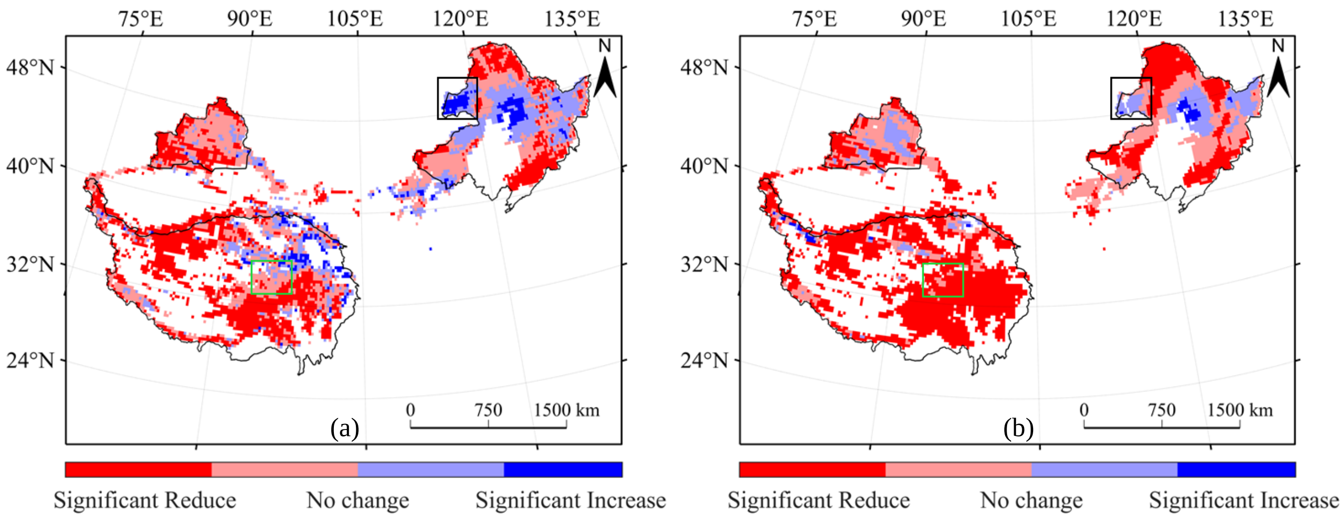

Figure 11. Trend analysis of snow depth during the period 1987-2018: (a) RF product; (b) WESTDC data. Light red and light blue represent no significant trend changes.

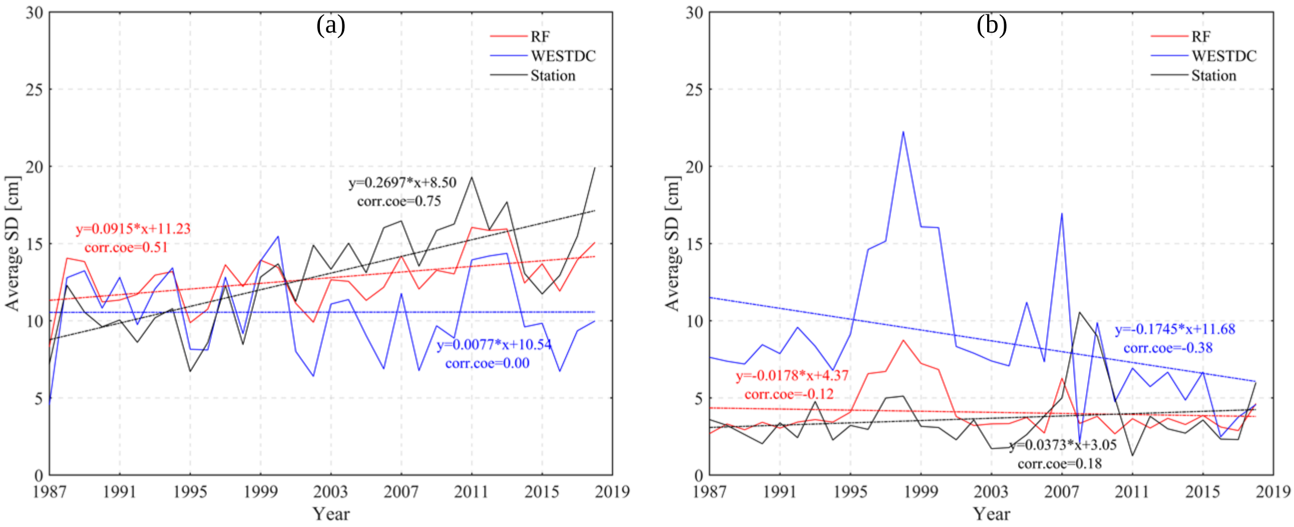

Figure 12. Comparison of changing trends of snow depth between RF estimates and WESTDC product in specific areas of (a) NE and (b) QTP.