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.

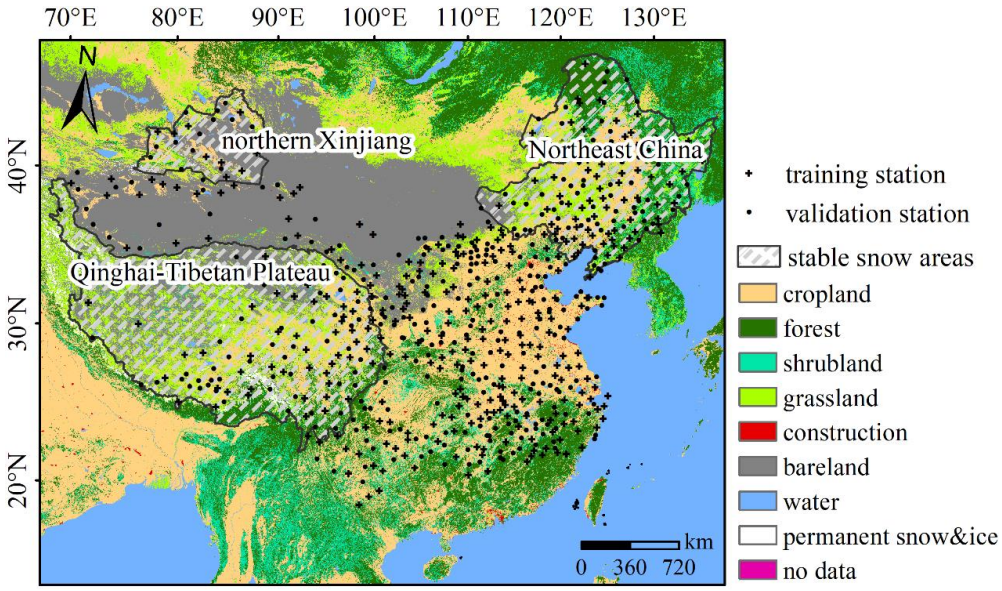

Figure 1. Spatial distribution of the weather stations and land cover types in the study area. There are three stable snow cover areas in China: Northeast China (NE), northern Xinjiang (XJ) and the Qinghai-Tibetan Plateau (QTP).

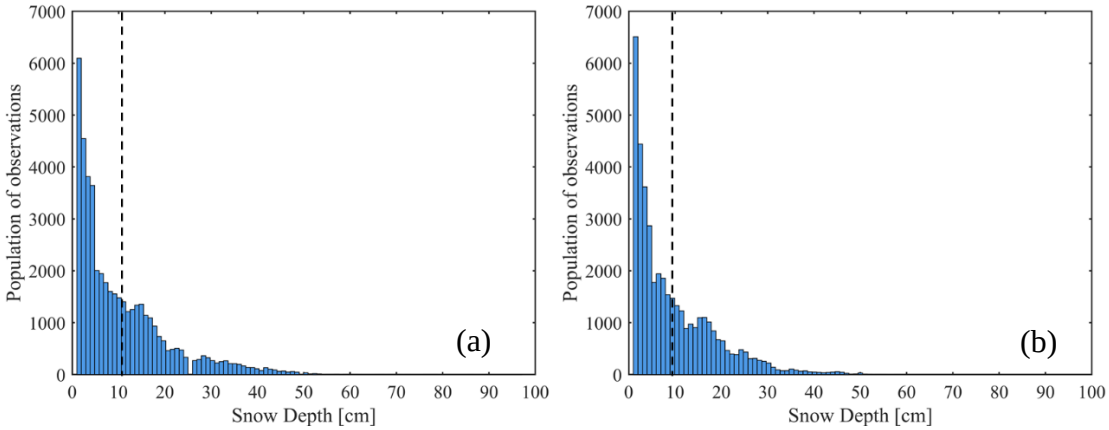

Figure 2. Histograms of snow depth observations from (a) training and (b) validation stations. The average values (black dashed lines) are equal to 10.5 cm and 9.8 cm, respectively.

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.

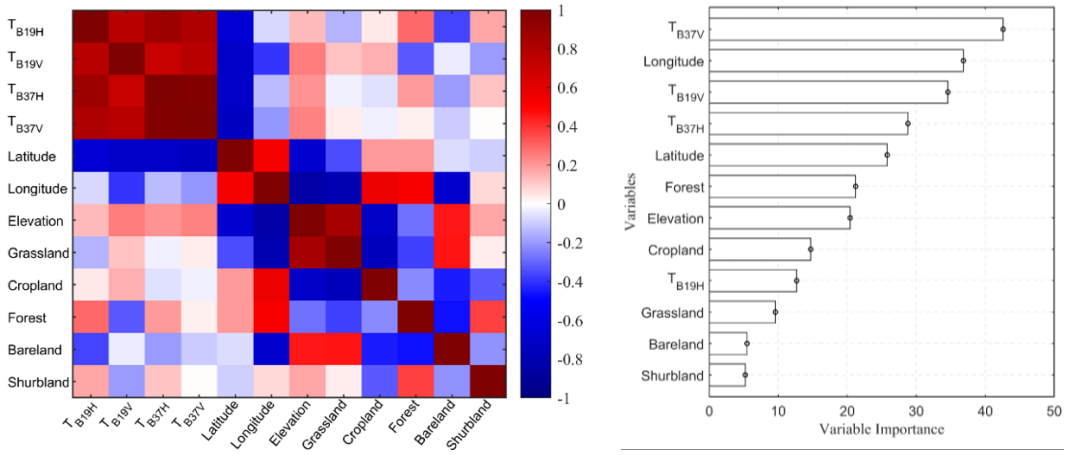

Figure 3. Correlations between the predictor variables (left) and the ranking of variable importance (right). The importance of variables, referred to as Mean Decrease Accuracy (MDA) in RF model, is obtained by averaging the difference in out-of-bag error estimation before and after the permutation over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

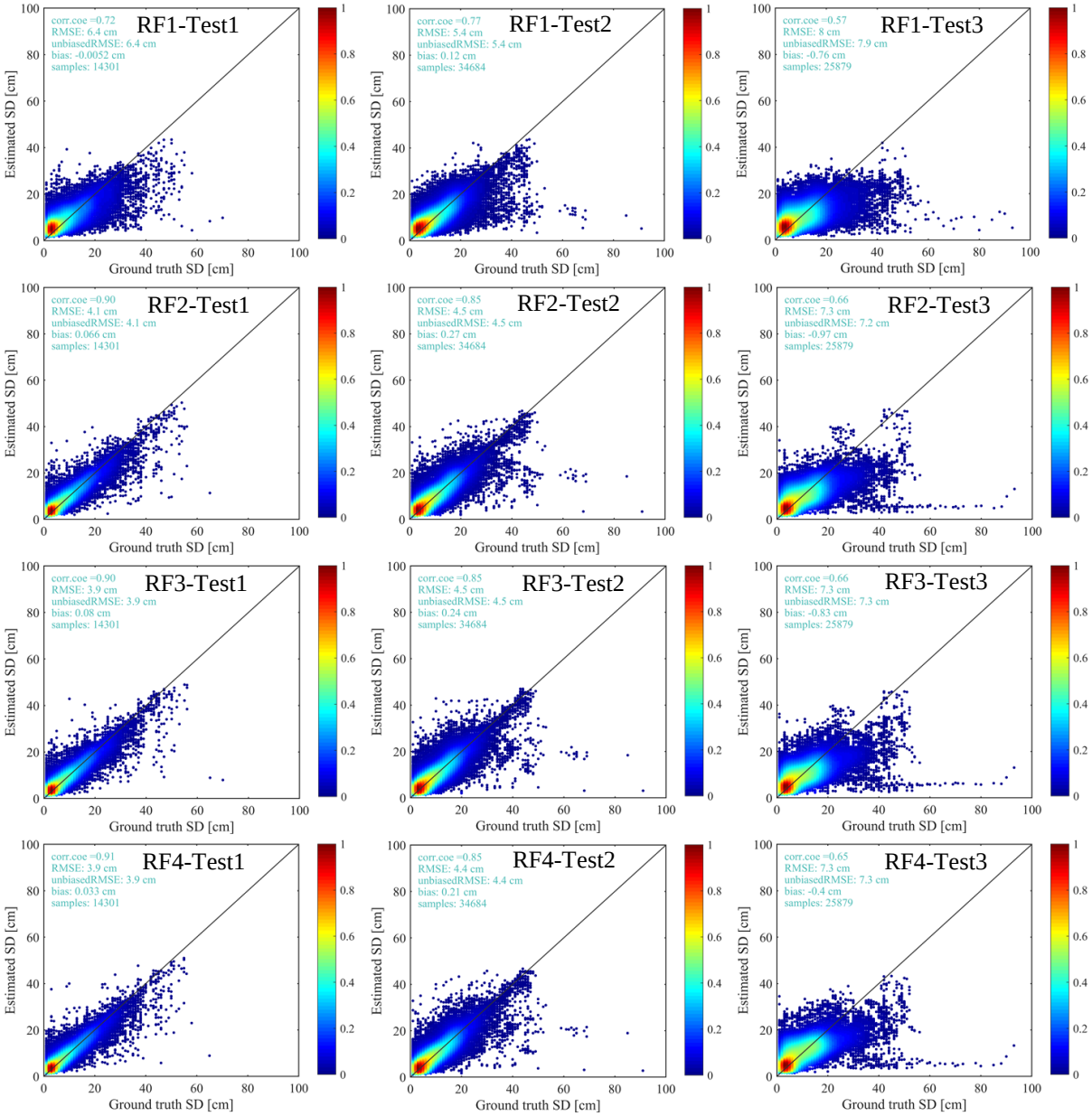

Figure 4. The color-density scatterplots of the estimated snow depth with four fitted RF algorithms and the ground truth snow depth. The four trained RF algorithms (RF1, RF2, RF3, RF4) were evaluated with three validation datasets (Test1, Test2, Test3).

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).

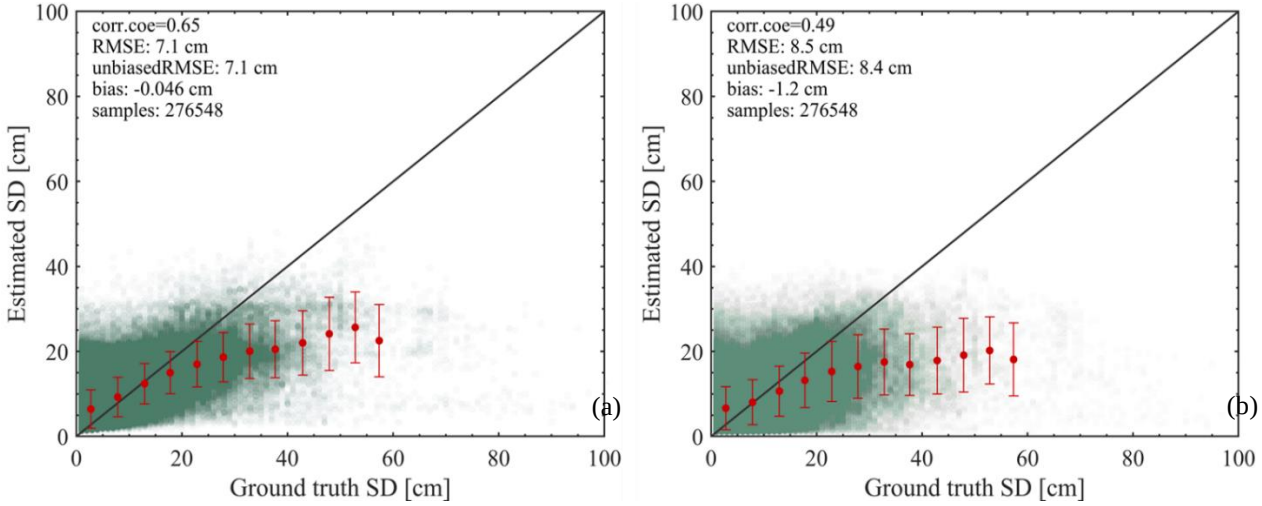

Figure 5. Scatterplots of the estimated snow depth and the ground truth observation for (a) RF and (b) WESTDC products.

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.

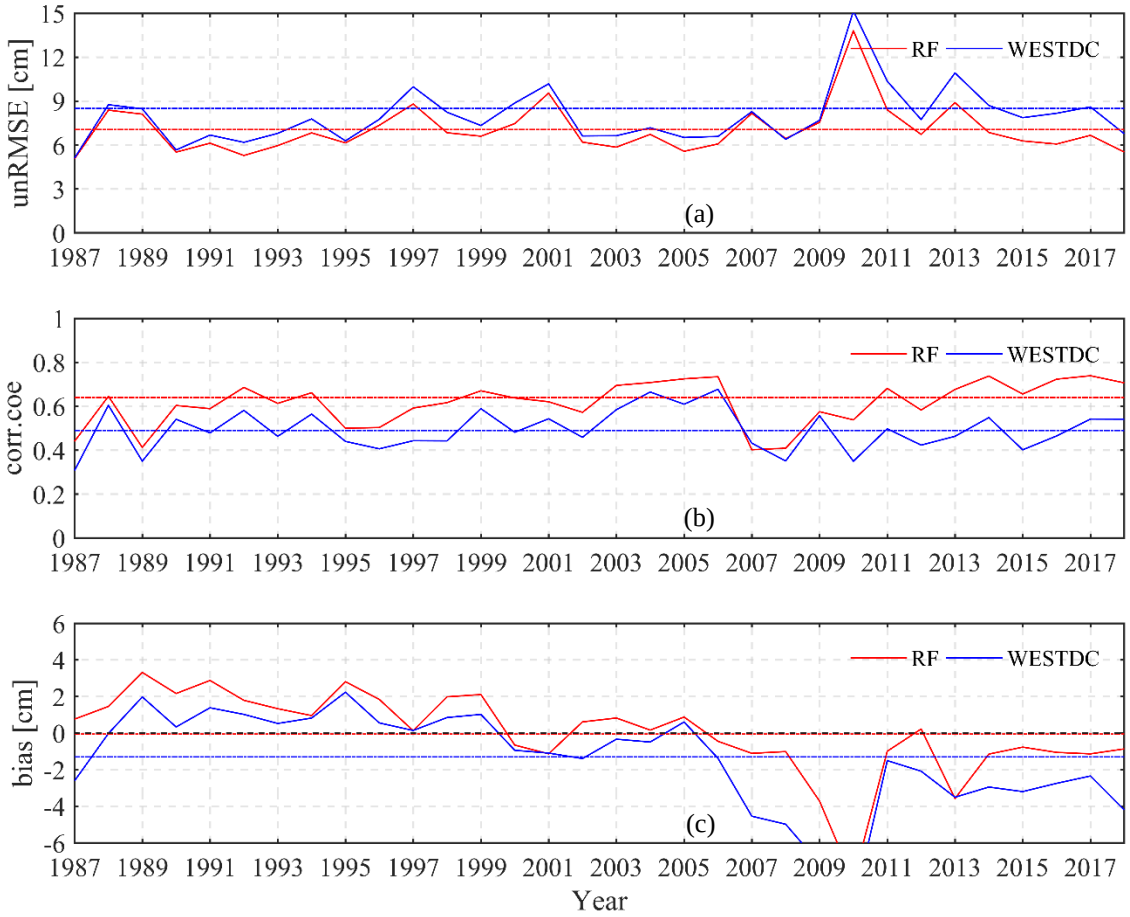

Figure 6. Time series of (a) unbiased RMSE (unRMSE), (b) correlation coefficient (corr.coe) and (c) bias for RF and WESTDC products. The colorful dashed lines represent mean values of assessment indexes.

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).

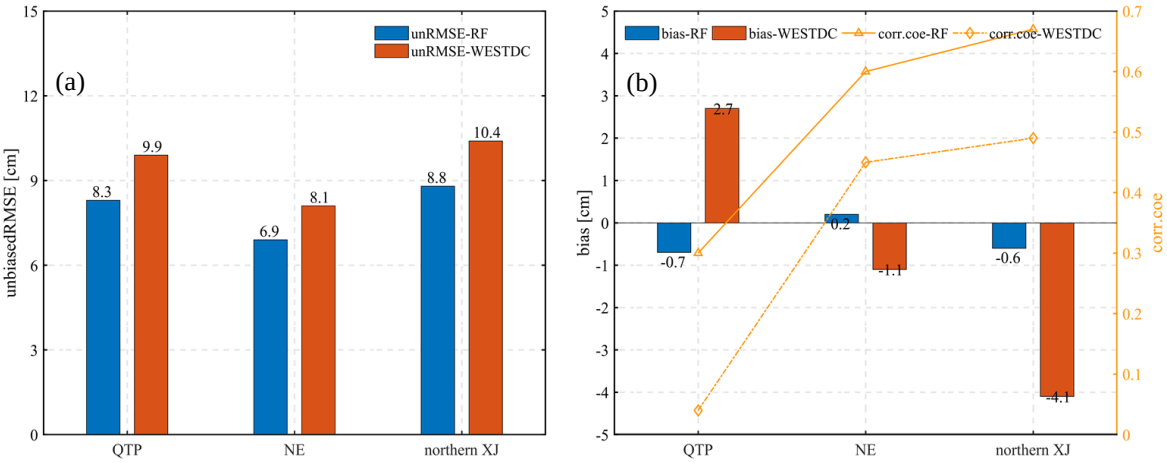

Figure 7. The validation of RF and WESTDC snow depth products in the three stable snow cover areas in China with respect to (a) the unbiased RMSE, (b) bias and correlation coefficient.

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."