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).

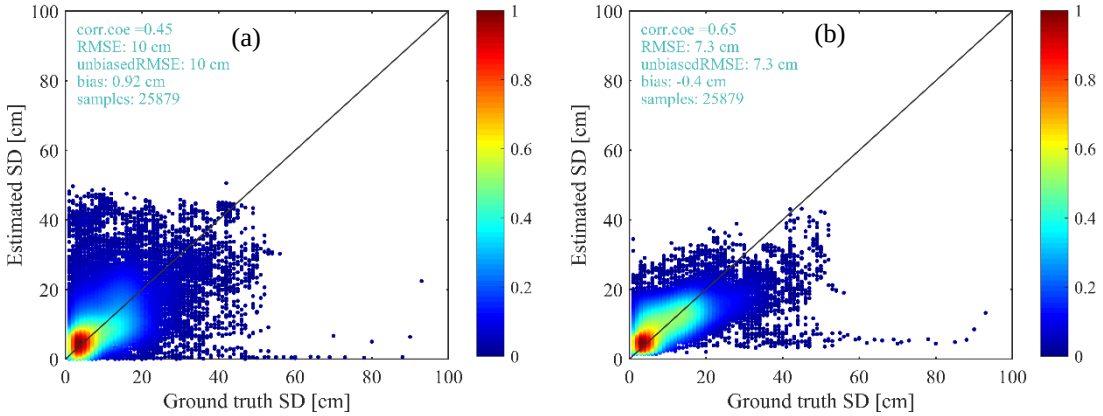

Figure 1. Comparison between (a) ANN and (b) RF with respect to snow depth estimation in China.

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.

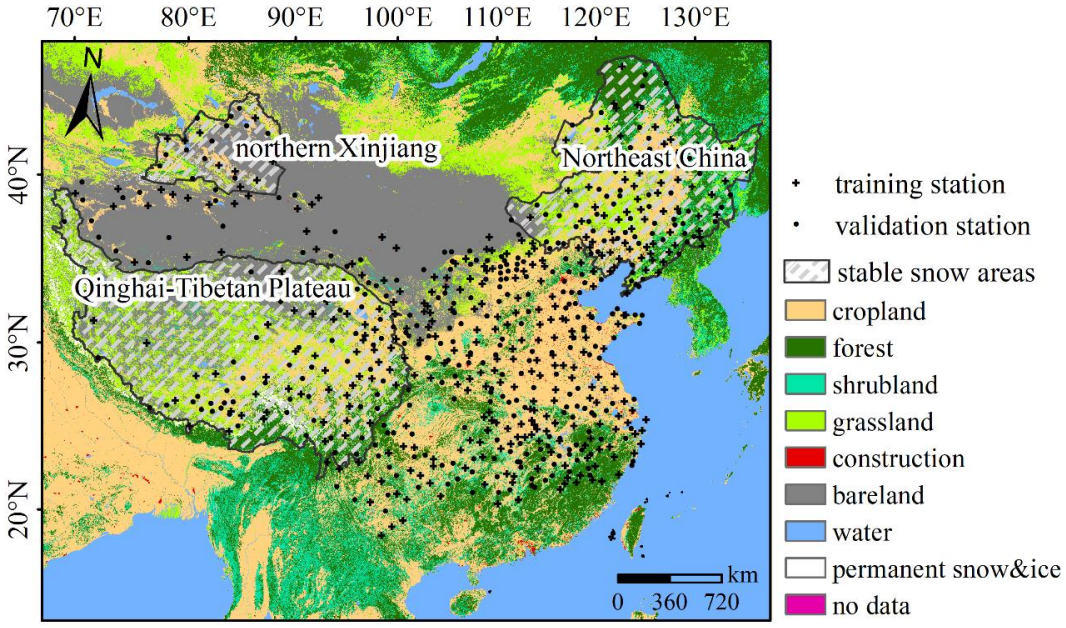

Figure 2. Spatial distribution of the weather stations and land cover types in the study area. There are three stable snow cover areas in China: Northeast China (NE), northern Xinjiang (XJ) and the Qinghai-Tibetan Plateau (QTP).

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).

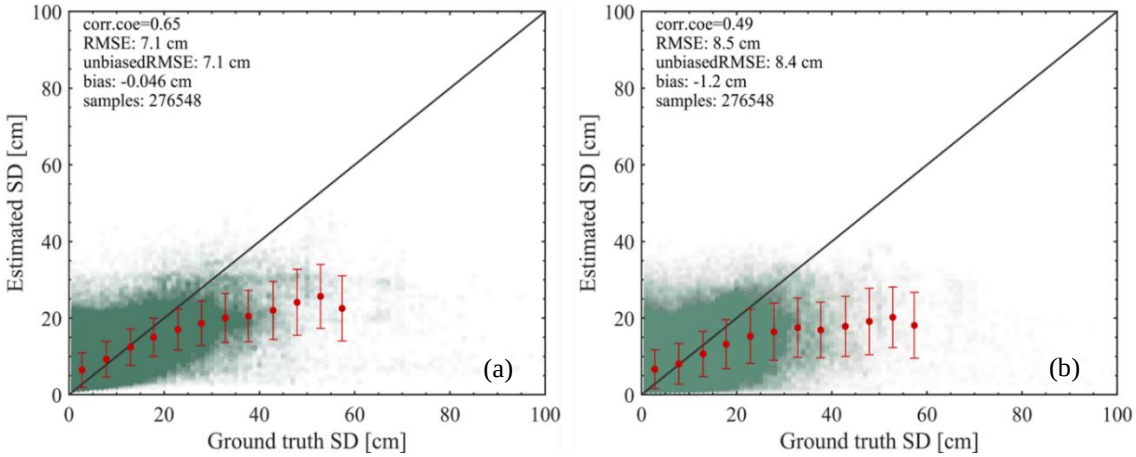

Figure 3. Scatterplots of the estimated snow depth and the ground truth observation for (a) RF and (b) WESTDC products.

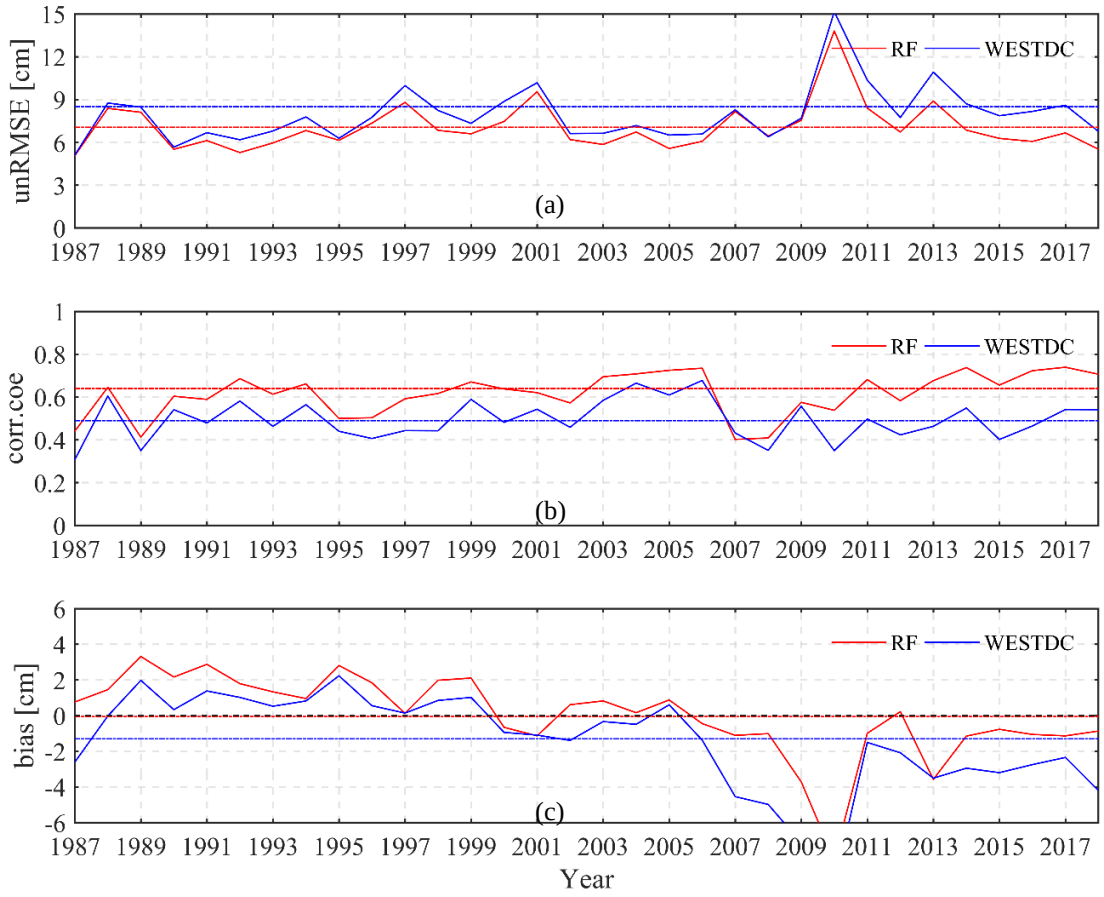

Figure 4. Time series of (a) unbiased RMSE (unRMSE), (b) correlation coefficient (corr.coe) and (c) bias for RF and WESTDC products. The colorful dashed lines represent mean values of assessment indexes.

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.

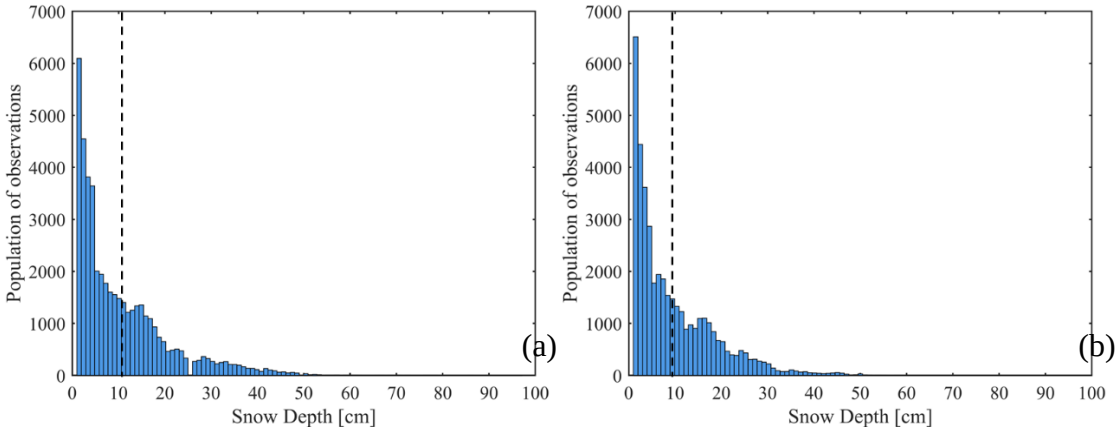

Figure 5. Histograms of snow depth observations from (a) training and (b) validation stations. The average values (black dashed lines) are equal to 10.5 cm and 9.8 cm, respectively.

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).

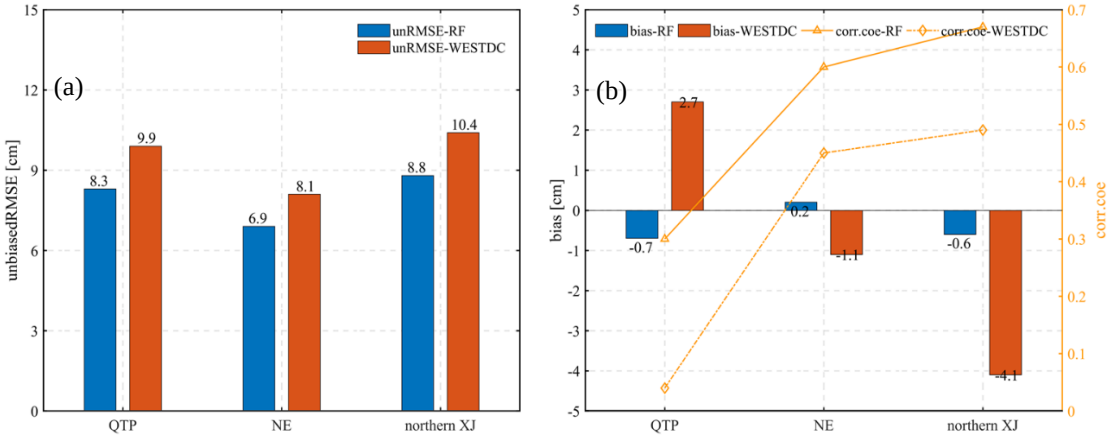

Figure 6. The validation of RF and WESTDC snow depth products in three stable snow cover areas in China with respect to (a) the unbiased RMSE, (b) bias and correlation coefficient.

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.

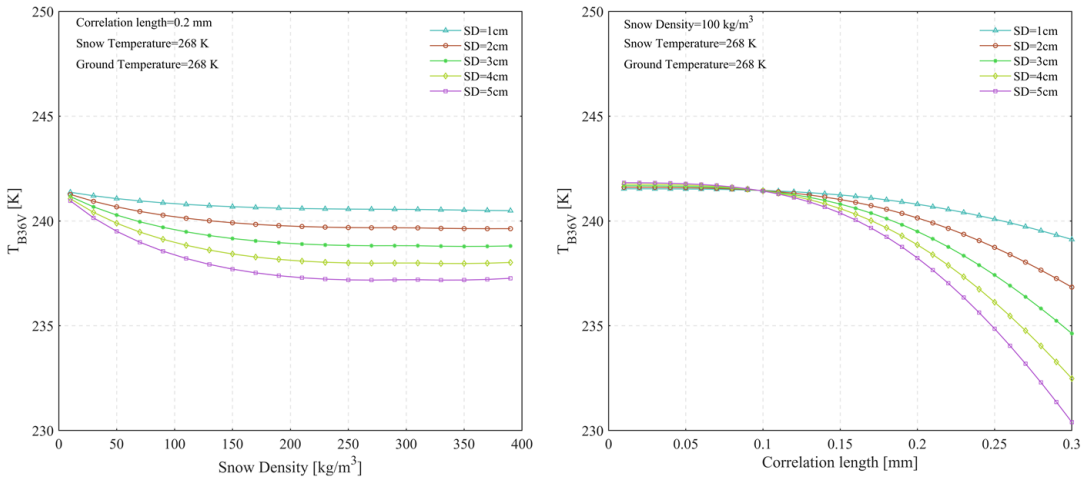

Figure 7. The sensitivity of snowpack stratigraphy to the passive microwave brightness temperature simulated with the MEMLS model.

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.

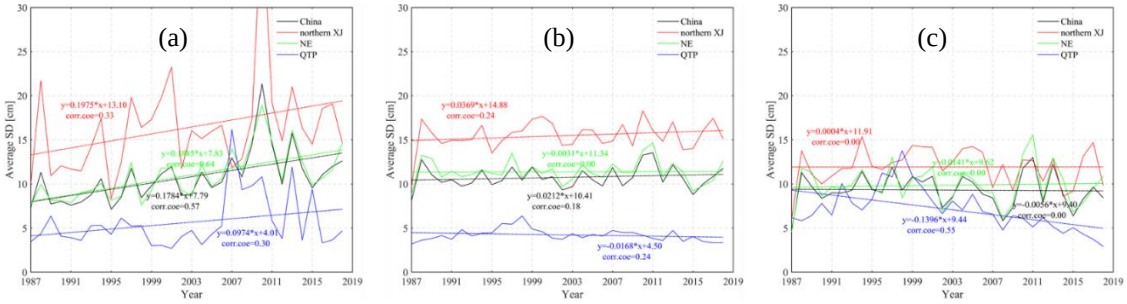

Figure 8. The trend analysis of snow depth based on (a) station observations, (b) RF estimates, and (d) WESTDC product in three stable snow cover areas in China. The correlation is statistically significant at the 0.05 level.

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.