**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.

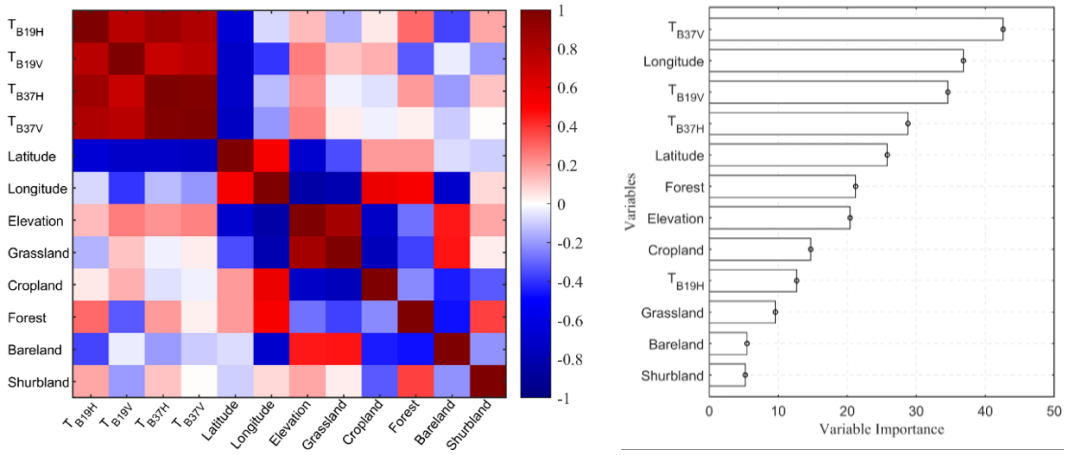


Figure 1. Correlations between the predictor variables (left) and the ranking of variable importance (right). The importance of variables, referred to as Mean Decrease Accuracy (MDA) in RF model, is obtained by averaging the difference in out-of-bag error estimation before and after the permutation over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

| 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 | | |

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 |

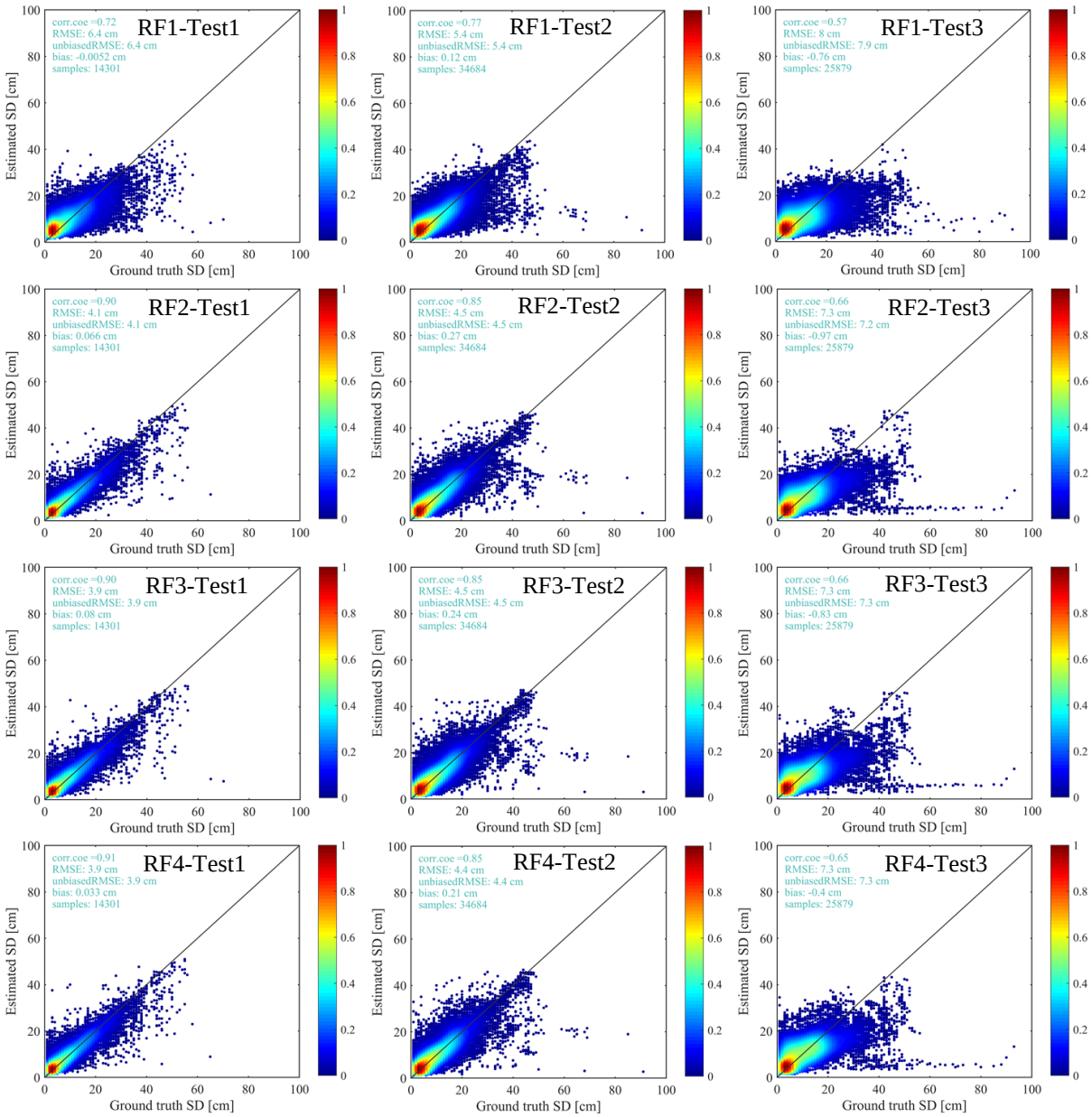


Figure 2. The color-density scatterplots of the estimated snow depth with four fitted RF algorithms and the ground truth snow depth. The four trained RF algorithms (RF1, RF2, RF3, RF4) were evaluated with three validation datasets (Test1, Test2, Test3).

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).

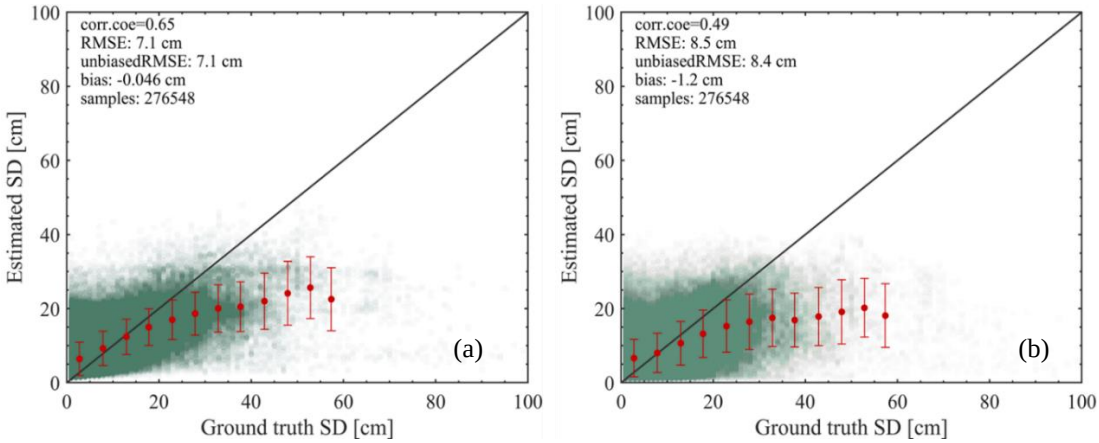


Figure 3. Scatterplots of the estimated snow depth and the ground truth observation for (a) RF and (b) WESTDC products.

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.

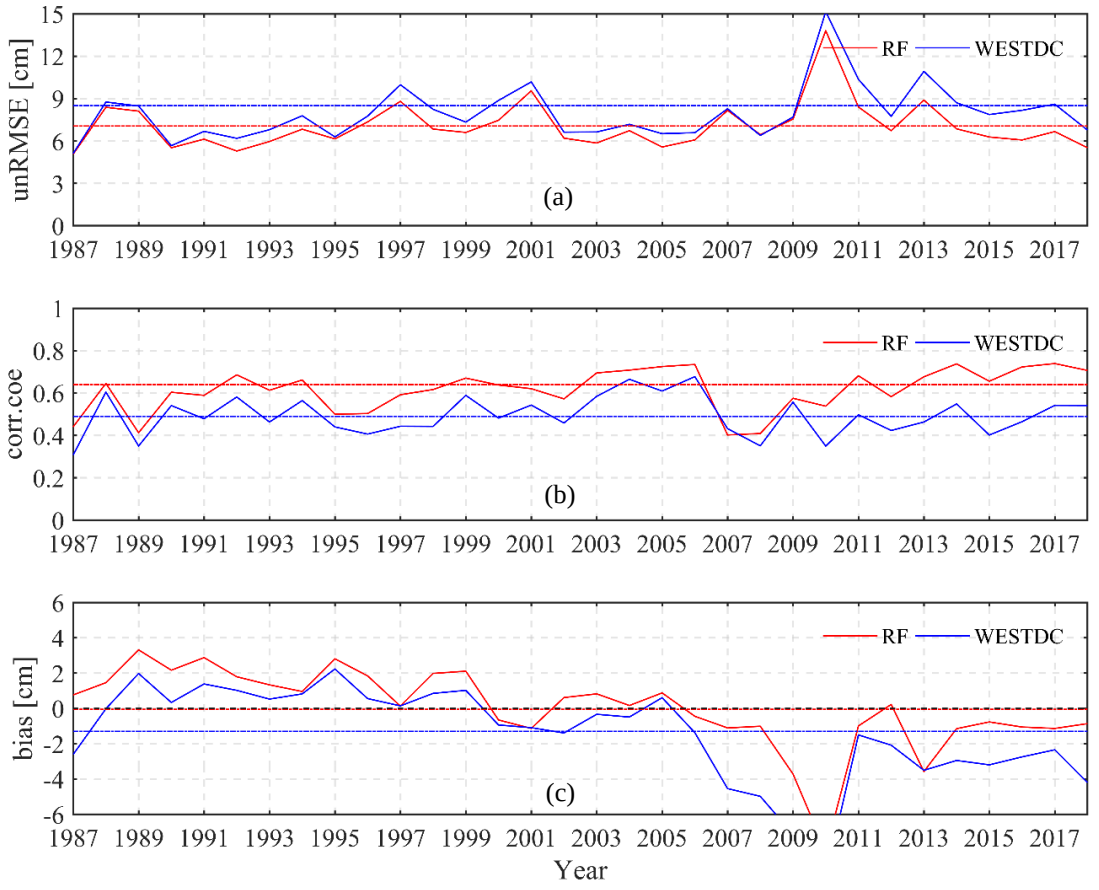


Figure 4. Time series of (a) unbiased RMSE (unRMSE), (b) correlation coefficient (corr.coe) and (c) bias for RF and WESTDC products. The colorful dashed lines represent mean values of assessment indexes.

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).

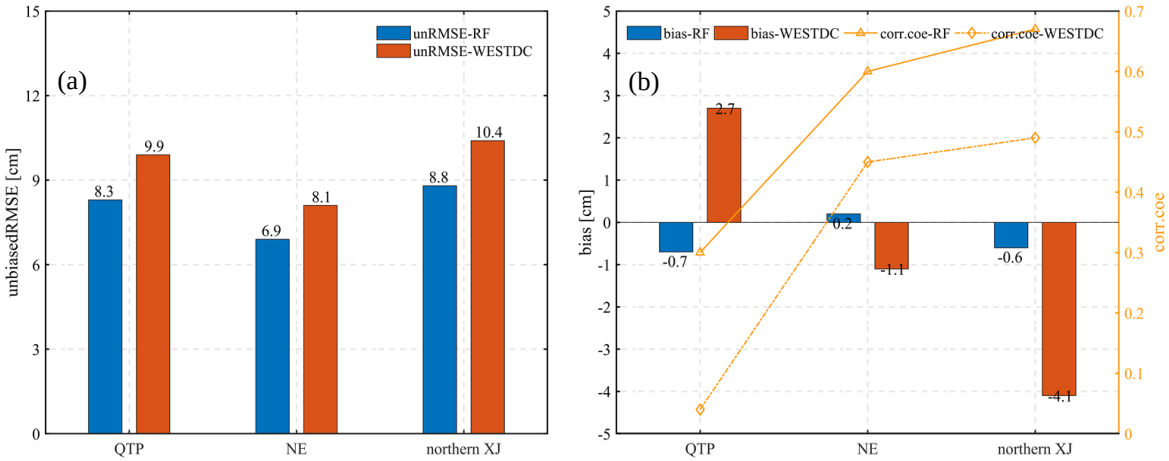


Figure 5. The validation of RF and WESTDC snow depth products in three stable snow cover areas in China with respect to (a) the unbiased RMSE, (b) bias and correlation coefficient.

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<br>D: 18:20 | A: 17:17<br>D: 05:17 | A: 17:58<br>D: 05:58 | A: 17:31<br>D: 05:31 |
| Frequency & footprint (GHz) : (km × km) | | 19.35: 45×68<br>23.235: 40×60<br>37: 24×36<br>85.5: 11×16 | | 19.35: 42×70<br>23.235: 42×70<br>37: 28×44<br>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.

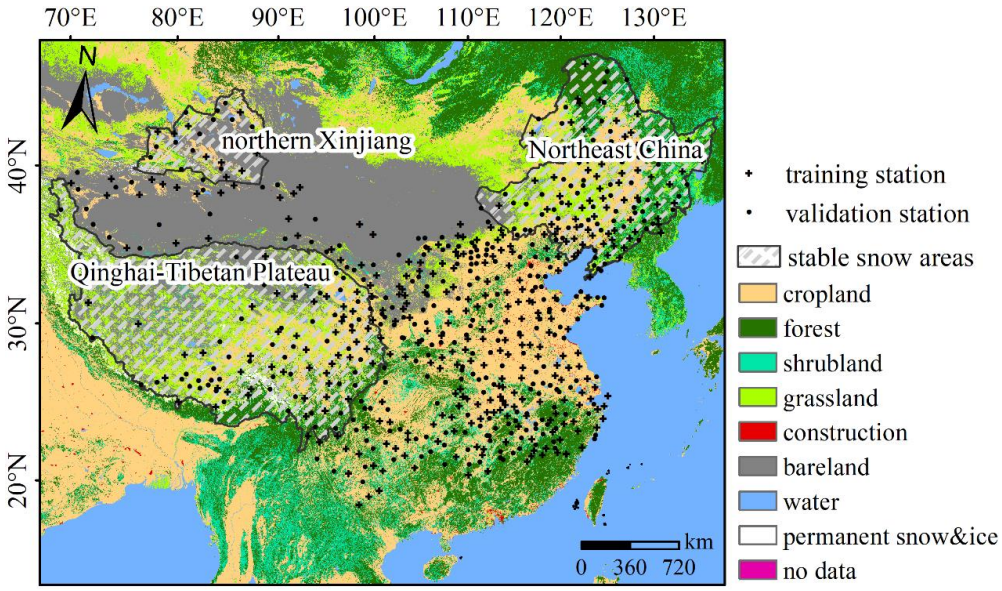


Figure 6. Spatial distribution of the weather stations and land cover types in the study area. There are three stable snow cover areas in China: Northeast China (NE), northern Xinjiang (XJ) and the Qinghai-Tibetan Plateau (QTP).

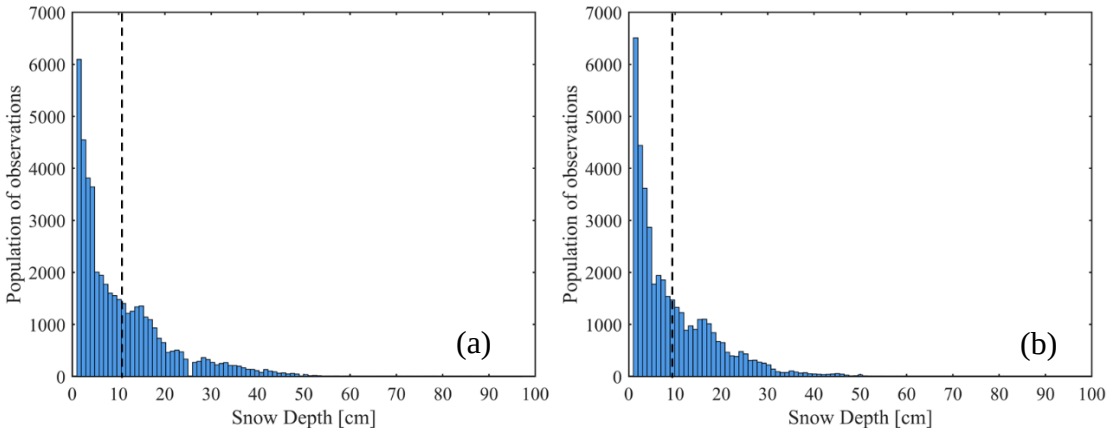


Figure 7. Histograms of snow depth observations from (a) training and (b) validation stations. The average values (black dashed lines) are equal to 10.5 cm and 9.8 cm, respectively.

**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).