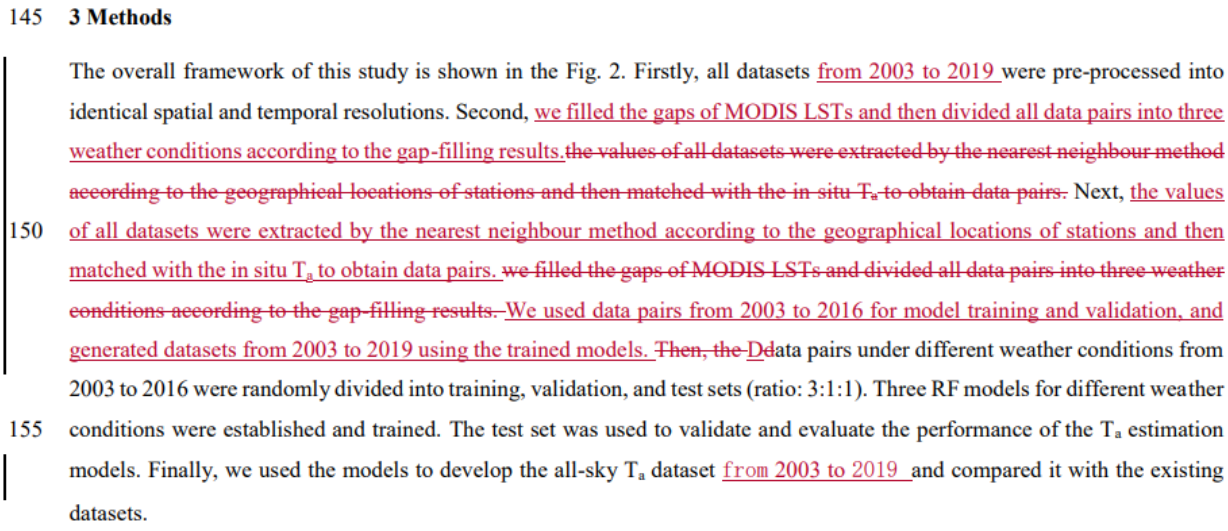
Response to Referee #1 Comments

We thank Referee #1 for the valuable and constructive comments on our manuscript. A point-by-point response to all comments is listed below.

Point 1: I'm wondering whether the data from 2003-2016 or 2003-2019 is used and produced. There seems to be inconsistency in the paper regarding the temporal period of the study.

Response 1: Thank you for your comments. We used data from 2003 to 2016 for model training and validation, and generated datasets from 2003 to 2019 using the trained models. Specifically, the data pairs from 2003 to 2016 were randomly divided into training, validation, and test sets (ratio: 3:1:1). Among them, training set was used for model training, validation set was used to determine the best model parameters, and test set was used to evaluate the final model performance. After model training, we used the models to develop the all-sky T_a dataset from 2003 to 2019. We have added the details on page 7, lines 146-156 in the revised manuscript:



3 Methods

The overall framework of this study is shown in the Fig. 2. Firstly, all datasets from 2003 to 2019 were pre-processed into identical spatial and temporal resolutions. Second, we filled the gaps of MODIS LSTs and then divided all data pairs into three weather conditions according to the gap-filling results.~~the values of all datasets were extracted by the nearest neighbour method according to the geographical locations of stations and then matched with the in situ T_a to obtain data pairs.~~ Next, the values of all datasets were extracted by the nearest neighbour method according to the geographical locations of stations and then matched with the in situ T_a to obtain data pairs. ~~we filled the gaps of MODIS LSTs and divided all data pairs into three weather conditions according to the gap-filling results.~~ We used data pairs from 2003 to 2016 for model training and validation, and generated datasets from 2003 to 2019 using the trained models. ~~Then, the~~ Ddata pairs under different weather conditions from 2003 to 2016 were randomly divided into training, validation, and test sets (ratio: 3:1:1). Three RF models for different weather conditions were established and trained. The test set was used to validate and evaluate the performance of the T_a estimation models. Finally, we used the models to develop the all-sky T_a dataset from 2003 to 2019 and compared it with the existing datasets.

Point 2: For vadiation of the study, how is the performance of the dataset/model if validation is carried out using a time period different from training period? For example, training is done using data from 2003 to 2016 and validation is done using data from 2017-2019? This is to see whether the training coeffients or RF models can be used after Terra/Aqua fail in the future.

Response 2: Thank you for your comments. We trained the models with the training set from 2003 to 2016, and further evaluated the models with data pairs from 2017 to 2019, which was not used for model training at all. The overall R^2, MAE, RMSE, and bias of the validation set were 0.982, 1.233 K, 1.611 K, and -0.340 K, respectively. The RMSE was slightly higher for the validation results using data from 2017 to 2019 compared to the validation results using the test set from 2003 to 2016 (1.611 K vs.

1.409 K). We found that there were certain differences in the T_a distribution between the two time periods. And the difference in the data distribution between the training set and the validation set may result in a slight decrease in the performance of the machine learning models on the validation set. Considering the data distribution range of T_a, we consider a difference of about 0.2 K to be acceptable. In general, the RF models have good generalization ability and can predict T_a of other years that have not been learned at all with satisfactory accuracy. We have added the content on page 19, lines 354-369 in the revised manuscript:

4.2 Independent validation

We further evaluated the models with data pairs from 2017 to 2019, which was not used for model training at all. The overall R^2, MAE, RMSE, and bias of the estimated all-sky T_a were 0.982, 1.233 K, 1.611 K, and -0.340 K, respectively. Figure 8 shows the density scatter plots of the estimated T_a against the in situ T_a from 2017 to 2019 under all weather conditions. The RMSE was slightly higher for the validation results using data from 2017 to 2019 compared to the validation results using the test set from 2003 to 2016 (1.611 K vs. 1.409 K). The histograms of T_a data distribution for 2003–2016 and 2017–2019 are shown in Fig. 9 (a) and (b), respectively. Obviously, there are differences in the data distribution of T_a between the two time periods. The T_a from 2003 to 2016 was slightly higher, with a more pronounced peak at around 295 K in the data distribution histogram. The difference in the data distribution between the training set and the validation set may result in a slight decrease in the performance of the machine learning models on the validation set. Considering the data distribution range of T_a, a difference of about 0.2 K in RMSE is acceptable. In general, after learning from a sufficiently large training set, the RF models have good generalization ability and can predict T_a of other years that have not been learned at all with good accuracy.

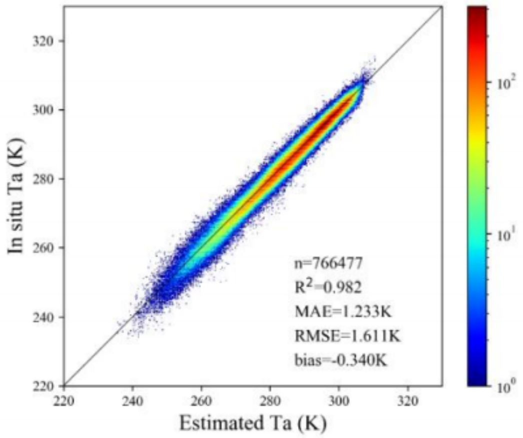


Figure 8. Density scatter plots of the estimated T_a and in situ T_a of independent validation results.

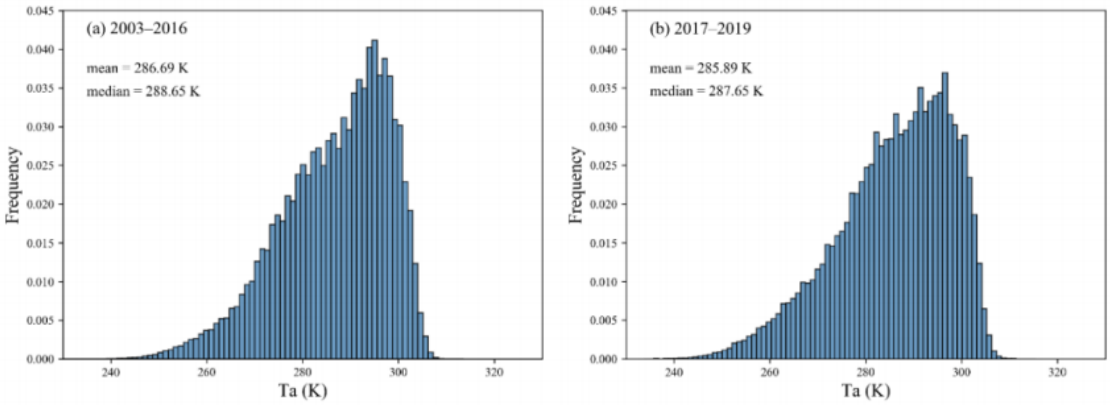


Figure 9. T_a data distribution for 2003–2016 (a) and 2017–2019 (b).

Point 3: I suggest to redo Figure 1 showing the number of data pairs and land types at these stations. You could use the color or the size of the symbol to provide such information.

Response 3: Thank you for your comments. We redid Figure 1 in the manuscript to show the spatial distribution and land cover types of the stations, as shown in Fig. 1 below. Each dot represents a station, and different colors correspond to different land cover types as shown in this figure legend. The land cover data used in the study is Finer Resolution Observation and Monitoring of Global Land Cover (FROM-GLC) version2 (2015_v1), which is a 30 m resolution global land cover maps (Gong et al., 2013). We have changed Figure 1 on page 6, lines 128-130 in the revised manuscript:

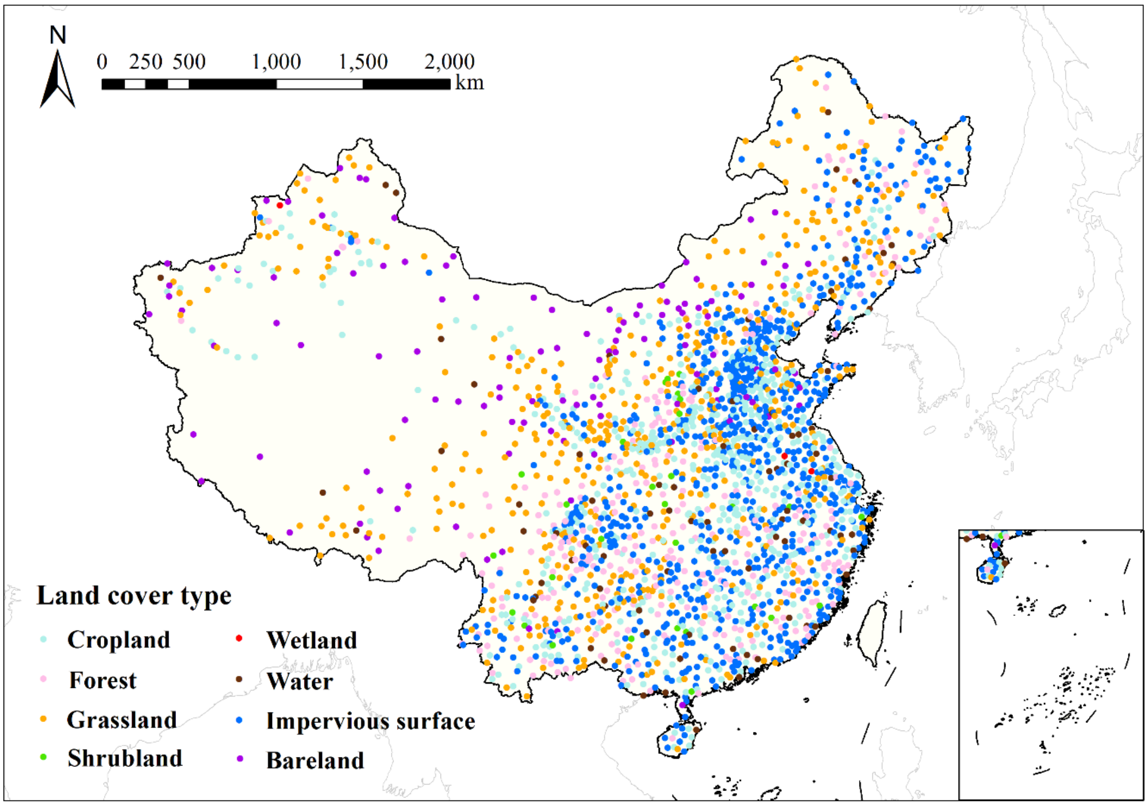


Figure 1. Study area and the location of meteorological stations used in this study. Each dot represents a station, and different colors correspond to different land cover types as shown in this figure legend.

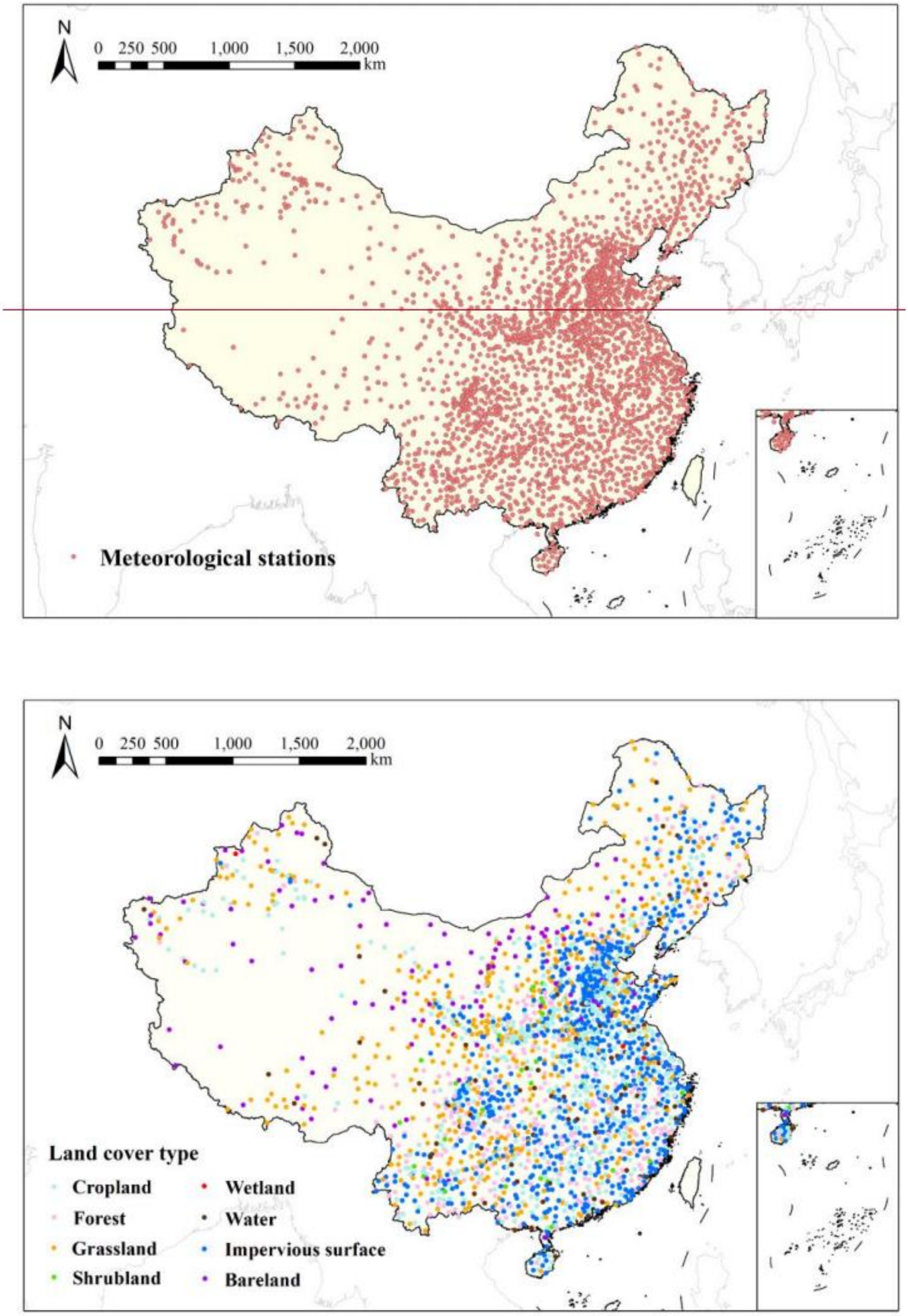


Figure 1. Study area and the location of meteorological stations ~~locations~~ used in this study. Each dot represents a station, and different colors correspond to different land cover types as shown in this figure legend.

We also calculated the number of data pairs from 2003 to 2016 for each station. Figure 2 below shows the number of data pairs of meteorological stations. Because station measurement data or satellite data or assimilation data were missing at some stations on some days, not all stations have data pairs equal to the total number of days. All 2384 meteorological stations used in this study have data pairs ranging from 1091 to 5113 over a 14-year period from 2003 to 2016. There were 2290 stations with data pairs greater than 5000, and only 6 stations with data pairs less than 3000. Overall, there is little difference in the number of data pairs at the station. Further combined with the analysis of the spatial distribution of model accuracy in Section 4 of the manuscript, it is concluded that the number of data pairs has no significant effect on the accuracy of T_a estimation.

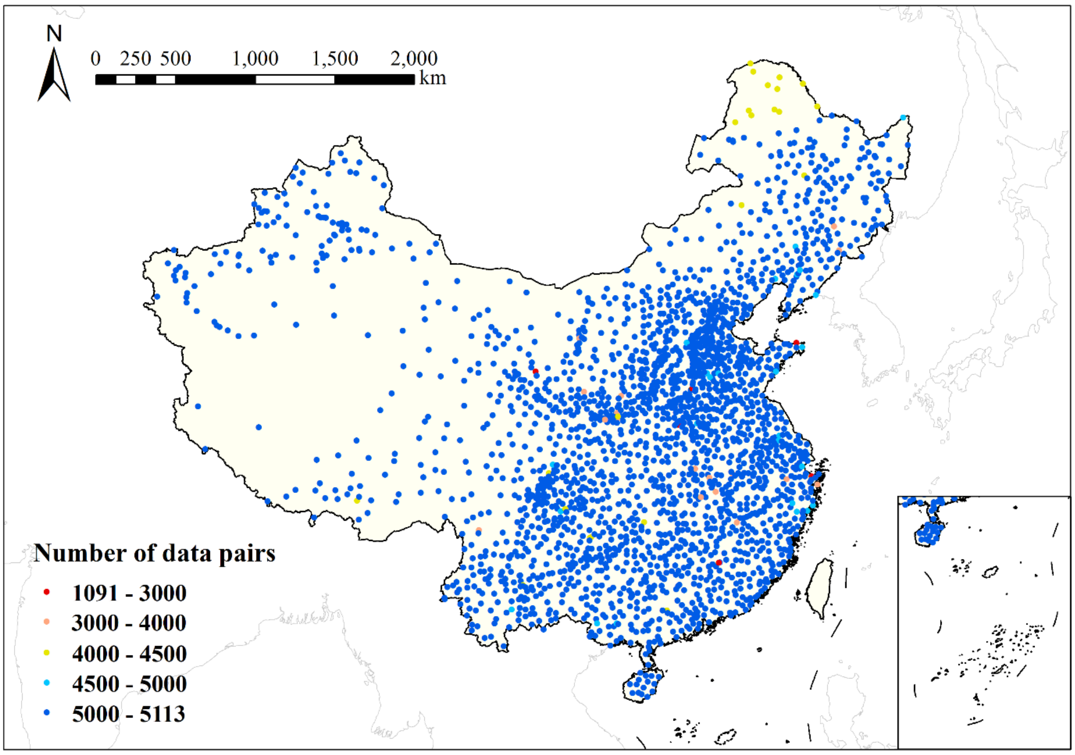


Figure 2. The spatial distribution of the number of data pairs from 2003 to 2016 of meteorological stations.

Point 4: Could you show the accuracy of the results as a joint function of surface types and surface temperature?

Response 4: Thank you for your comments. The relationship between land surface temperature (LST) and error under 8 surface types is represented by different colors as shown in the legend in Fig. 3. The abscissa is the average of the four daily LSTs for a data pair, and the ordinate is the error, which is the difference between the estimated T_a and the station measured T_a.

As can be seen from Fig. 3, for different surface types, the number of data pairs and the range of LST are different. The error range is also different. For each surface type, the errors showed no significant difference at different LST, and all present a normal distribution centered on 0 K. Therefore, the model performance varies with the surface types to some extent, but the estimation accuracy has no significant joint correlation with surface types and LST.

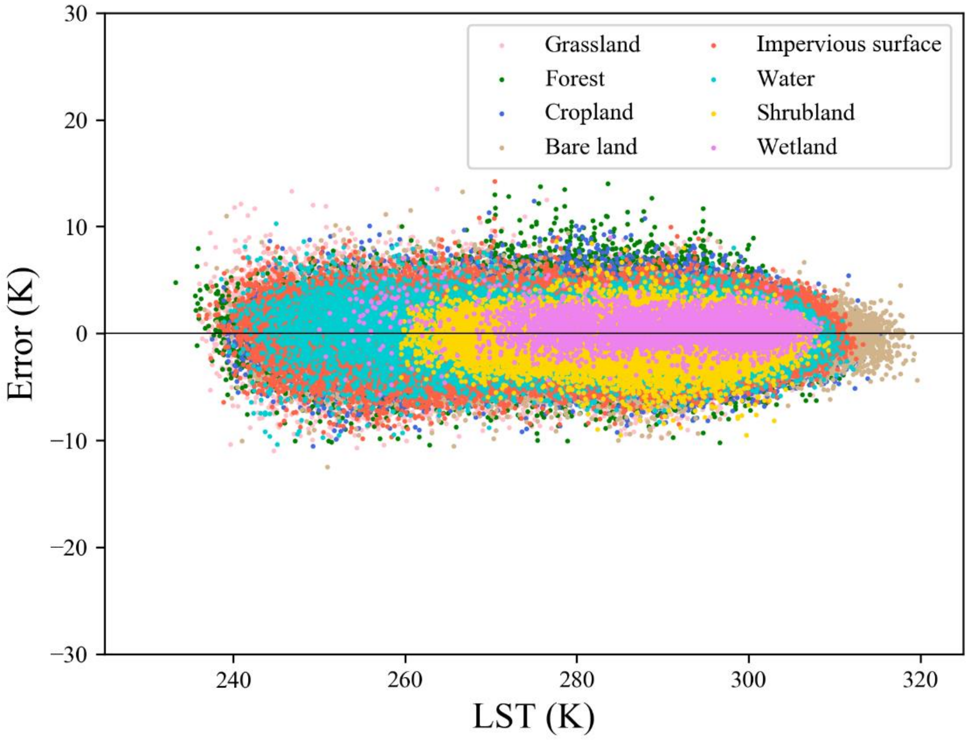


Figure 3. The relationship between LST and error under different surface types.

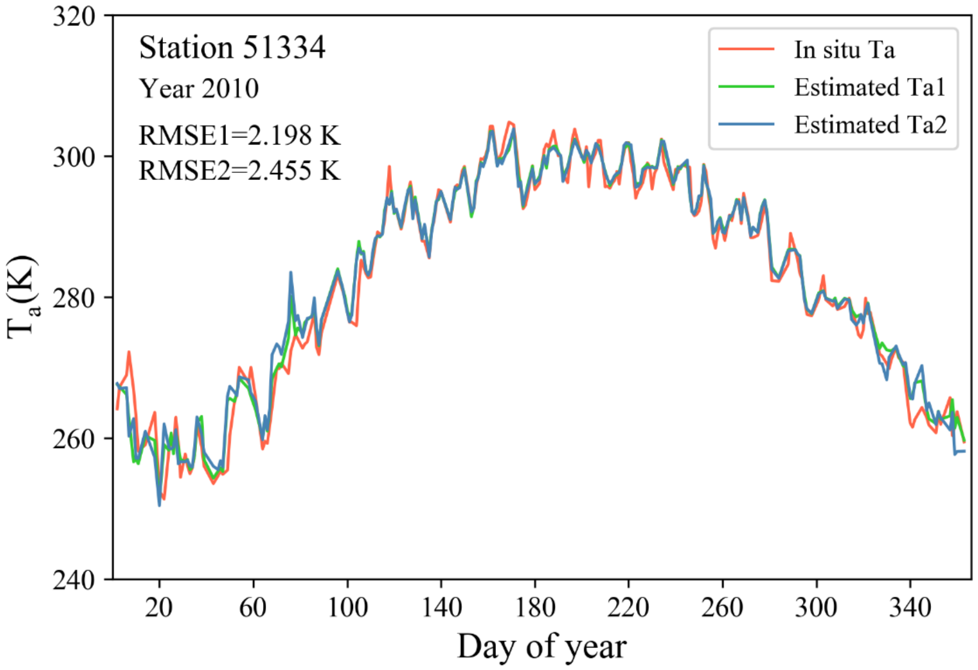
Point 5: If the FI factors are small for surface radiation measurements, why not remove them from your model?

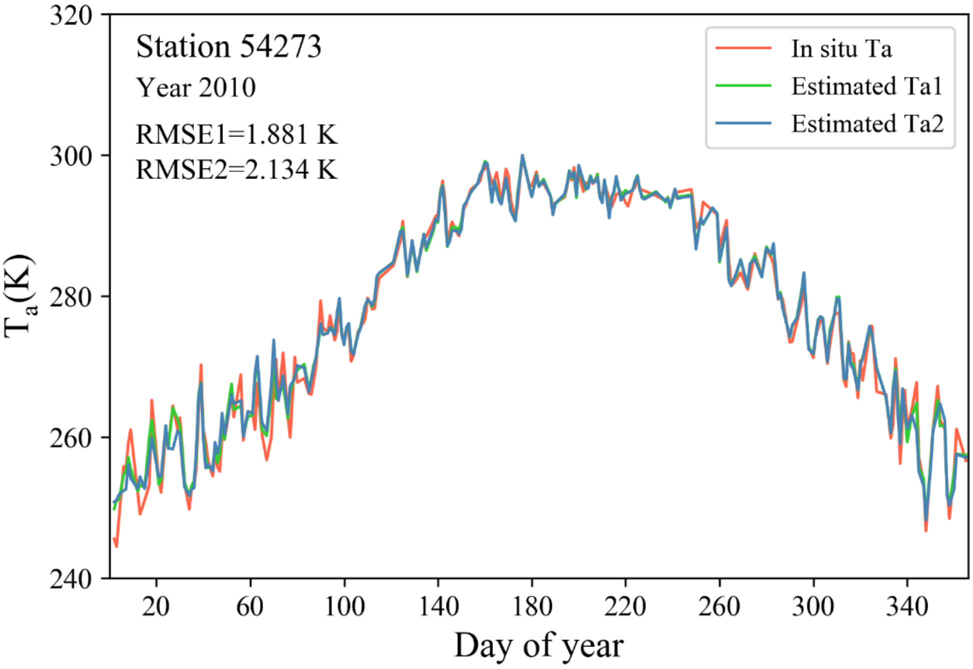
Response 5: Thank you for your comments. The radiation features help to reflect the heat exchange process between the surface and the atmosphere. In our experiment, we found that the FI factors of radiation features were small for the T_a estimation models. Table 1 lists the validation results for models with and without radiation features. It can be seen that, after removing DSR and ALB features, the overall RMSE values of the validation set for the three models increased by 0.02-0.06 K. Therefore, the radiation features have little influence on the overall accuracy of the models.

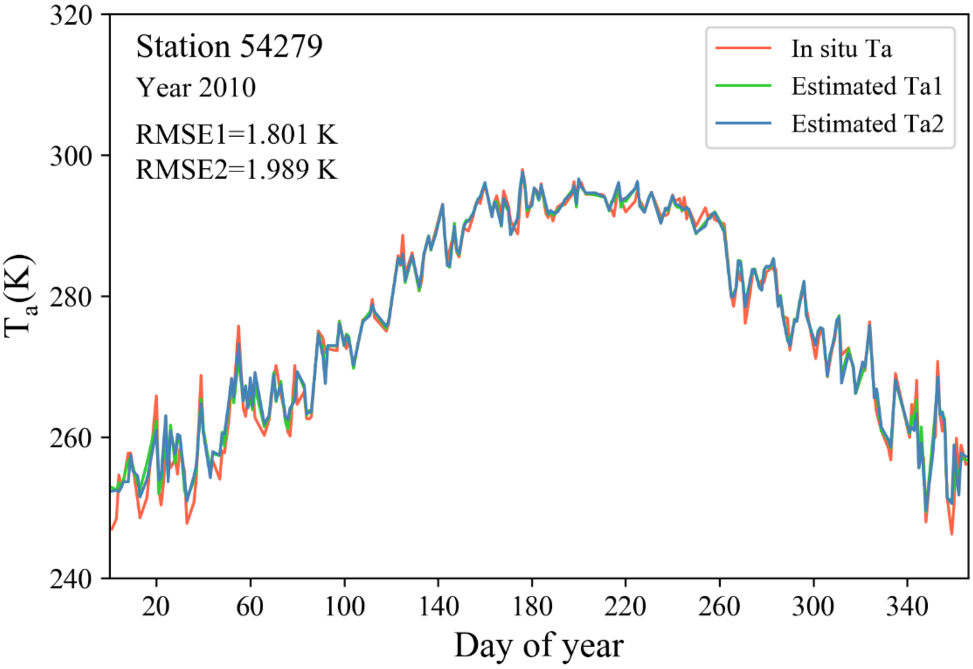
However, in the analysis of the results of some stations, it is found that the accuracy of the models including radiation features was higher than that of the models excluding radiation features at some stations. For example, Fig. 4 below shows the T_a annual curves of four stations in 2010. In the figure, the orange lines are the station measured T_a, while the green and blue lines are the T_a predicted by the models with and without radiation features, respectively. RMSE1 and RMSE2 are RMSE values for models with and without radiation features, respectively. The results showed that on some days, adding radiation features to the models helped improve the T_a estimation accuracy at some stations. Although there may be other collinear features in the models that make the information provided by them redundant, the radiation features can play a supplementary role in the case of some other features that do not perform well. Therefore, we finally decided to retain the radiation features in the T_a estimation models.

Table 1. Validation results for models with and without radiation features.

Model	Include radiation features		Not include radiation features	
	R^2	RMSE (K)	R^2	RMSE (K)
Clear-sky model	0.986	1.342	0.985	1.365
Cloudy-sky model I	0.984	1.440	0.984	1.468
Cloudy-sky model II	0.984	1.396	0.983	1.451
All	0.985	1.409	0.984	1.448




Station 54273
Year 2010
RMSE1=1.881 K
RMSE2=2.134 K
In situ Ta
Estimated Ta1
Estimated Ta2
T_a(K)
320
300
280
260
240
20
60
100
140
180
220
260
300
340
Day of year


Station 54279
Year 2010
RMSE1=1.801 K
RMSE2=1.989 K
In situ Ta
Estimated Ta1
Estimated Ta2
T_a(K)
320
300
280
260
240
20
60
100
140
180
220
260
300
340
Day of year

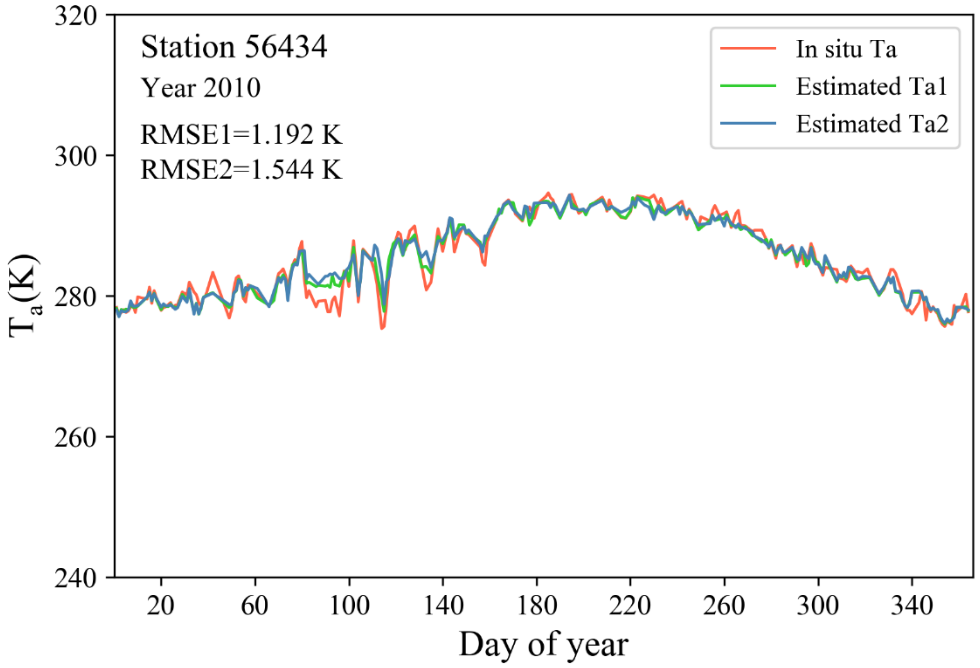
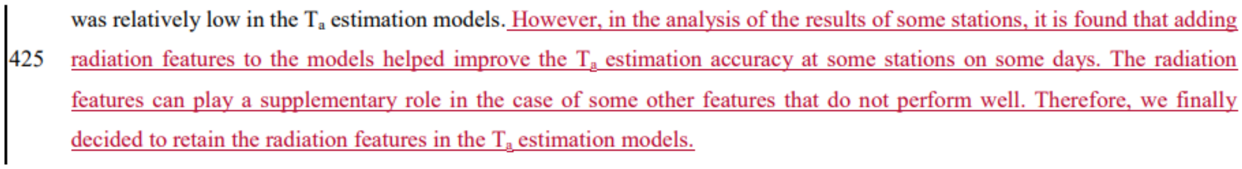


Figure 4. T_a annual curves of station 51334, station 54273, station 54279, and station 56434 in 2010. The orange lines are the station measured T_a, while the green and blue lines are the T_a predicted by the models with and without radiation features, respectively. RMSE1 and RMSE2 are RMSE values for models with and without radiation features, respectively.

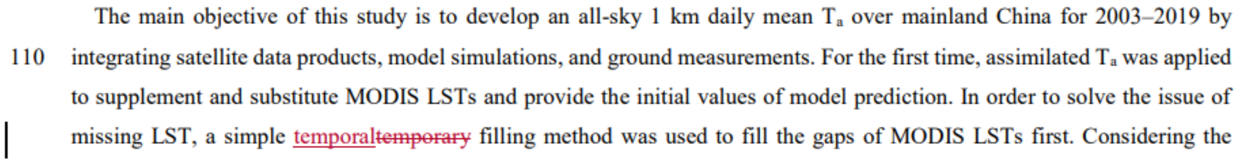
We have added the reason for retaining the radiation features on page 23, lines 424-427 in the revised manuscript:



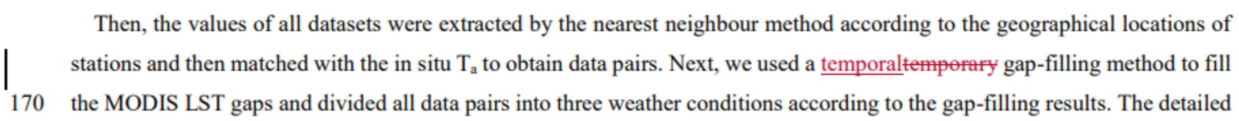
was relatively low in the T_a estimation models. However, in the analysis of the results of some stations, it is found that adding radiation features to the models helped improve the T_a estimation accuracy at some stations on some days. The radiation features can play a supplementary role in the case of some other features that do not perform well. Therefore, we finally decided to retain the radiation features in the T_a estimation models.

Point 6: There are places in the paper using "temporary gap filling model", but it should be "temporal" instead of "temporary".

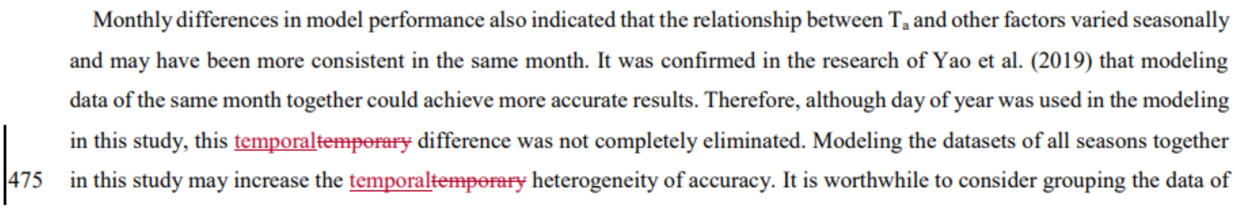
Response 6: Thank you for your comments. We have modified the words on page 4, line 112, and page 8, line 169 and page 26, lines 474-475, and page 33, line 543 in the revised manuscript:



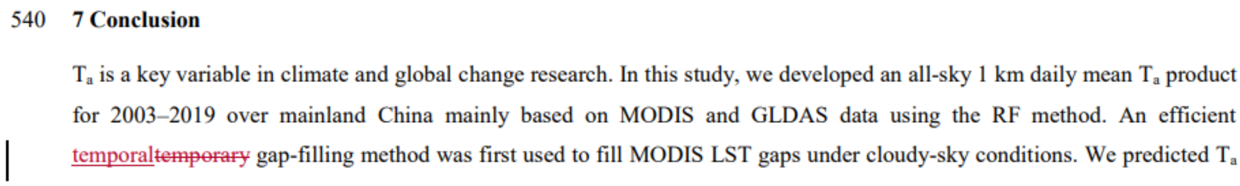
The main objective of this study is to develop an all-sky 1 km daily mean T_a over mainland China for 2003–2019 by integrating satellite data products, model simulations, and ground measurements. For the first time, assimilated T_a was applied to supplement and substitute MODIS LSTs and provide the initial values of model prediction. In order to solve the issue of missing LST, a simple temporal~~temporary~~ filling method was used to fill the gaps of MODIS LSTs first. Considering the



Then, the values of all datasets were extracted by the nearest neighbour method according to the geographical locations of stations and then matched with the in situ T_a to obtain data pairs. Next, we used a temporal~~temporary~~ gap-filling method to fill the MODIS LST gaps and divided all data pairs into three weather conditions according to the gap-filling results. The detailed



Monthly differences in model performance also indicated that the relationship between T_a and other factors varied seasonally and may have been more consistent in the same month. It was confirmed in the research of Yao et al. (2019) that modeling data of the same month together could achieve more accurate results. Therefore, although day of year was used in the modeling in this study, this temporal~~temporary~~ difference was not completely eliminated. Modeling the datasets of all seasons together in this study may increase the temporal~~temporary~~ heterogeneity of accuracy. It is worthwhile to consider grouping the data of



7 Conclusion

T_a is a key variable in climate and global change research. In this study, we developed an all-sky 1 km daily mean T_a product for 2003–2019 over mainland China mainly based on MODIS and GLDAS data using the RF method. An efficient temporal~~temporary~~ gap-filling method was first used to fill MODIS LST gaps under cloudy-sky conditions. We predicted T_a

Point 7: are the station Ta measurements used in the prediction of Ta?

Response 7: Thank you for your comments. In this study, the station T_a measurements were not used in the prediction of T_a, but were used in model training. The data pairs used for model training and validation consist of input features and station measured T_a at the stations. The input features of the models are LSTs, DSR, ALB, LAI, elevation, GLDAS T_a, day of year, latitude, and longitude. And the output variable is daily mean T_a.

Reference:

Gong, P., Wang, J., Yu, L., Zhao, Y., Zhao, Y., Liang, L., Niu, Z., Huang, X., Fu, H., and Liu, S.: Finer resolution observation and monitoring of global land cover: First mapping results with Landsat TM and ETM+ data, Int. J. Remote Sens., 34, 2607-2654, https://doi.org/10.1080/01431161.2012.748992, 2013.