Response to Referee #2 Comments

We thank Referee #2 for the valuable and constructive comments on our manuscript. A point-by-point response to all comments is listed below.

Point 1: My biggest concern is the statistic results may not be credible over southwest China. In southwest China, it rains in the most time of a year. The sunshine is rare that it is said "sunny weather seldom lasts for more than three days". In other words, the percentage of missing MODIS LST data is over 60% due to the presence of clouds. To control the uncertainty introduced by LST gap-filling, a temporal window of 2 days in this study was used to fill the gaps. Please provide detailed availability of LST for cloudy-sky model using this simple gap-filling method.

Response 1: Thank you for your comments. In this study, a simple multi-temporal method was used to fill the MODIS LST gaps. In order to balance the MODIS LST gap-filling rate with the large uncertainty caused by the large time threshold, we have conducted experiments with different time thresholds, and finally decided to set the time threshold of ± 2 days. The ratios of available values of four MODIS LSTs at all stations were 33.2 %, 37.6 %, 32.1 %, and 38.0 %, respectively, which increased to 73.0 %, 77.7 %, 72.4 %, and 77.3 %, respectively, after gap-filling.

Moreover, we counted the validation statistics of 485 stations located in southwest China, and the overall R^2 and RMSE were 0.978 and 1.428 K, respectively. The density scatter plot of the estimated T_a against the station observed T_a in southwest China under all weather conditions is shown in Fig. 1. The RMSE histogram of stations in southwest China is shown in Fig. 2, with a mean RMSE of 1.405 K. Of the 485 stations, 315 stations had RMSE values of less than 1.5 K, while only 9 stations had RMSE values of more than 2 K. Therefore, stations in southwest China have generally shown satisfactory performance, we consider this gap-filling method feasible for this study.

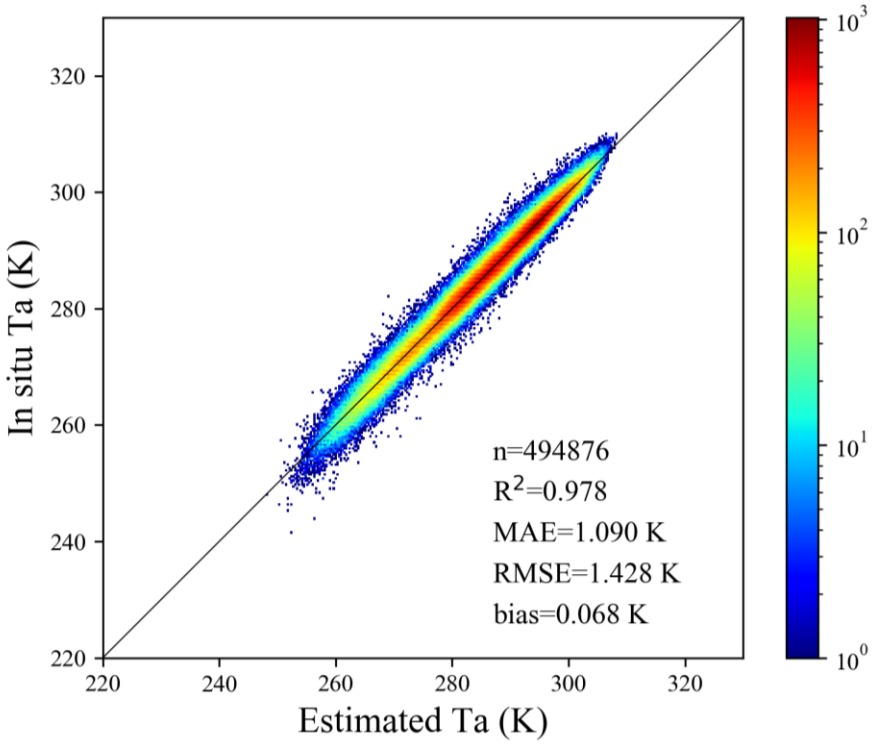


Figure 1. Density scatter plot of the estimated T_a against the station observed T_a in southwest China under all weather conditions.

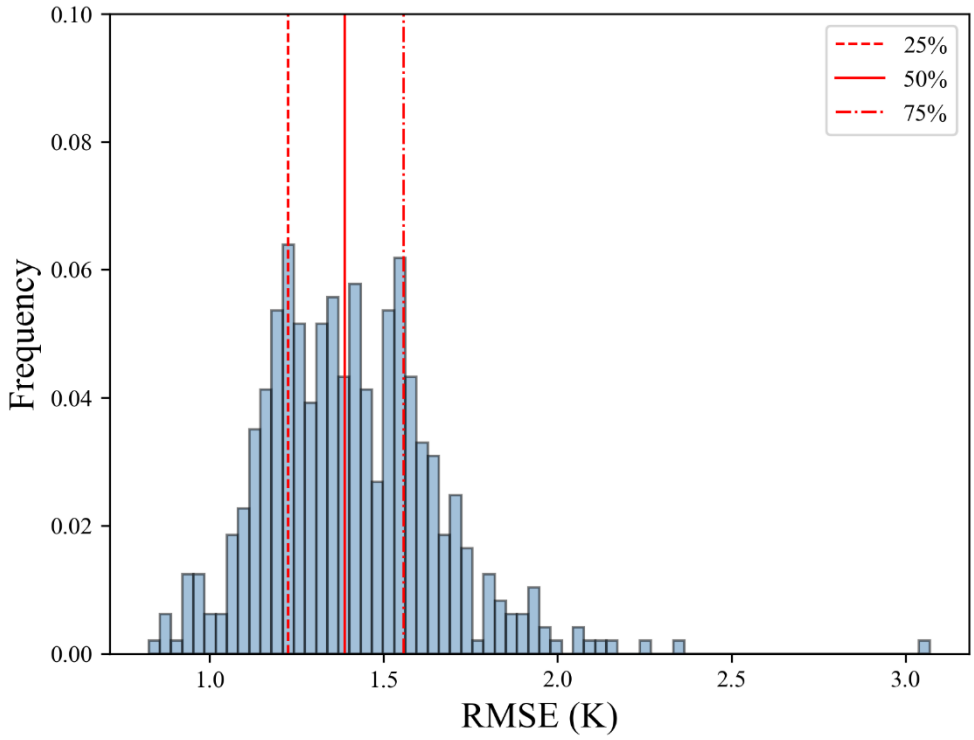
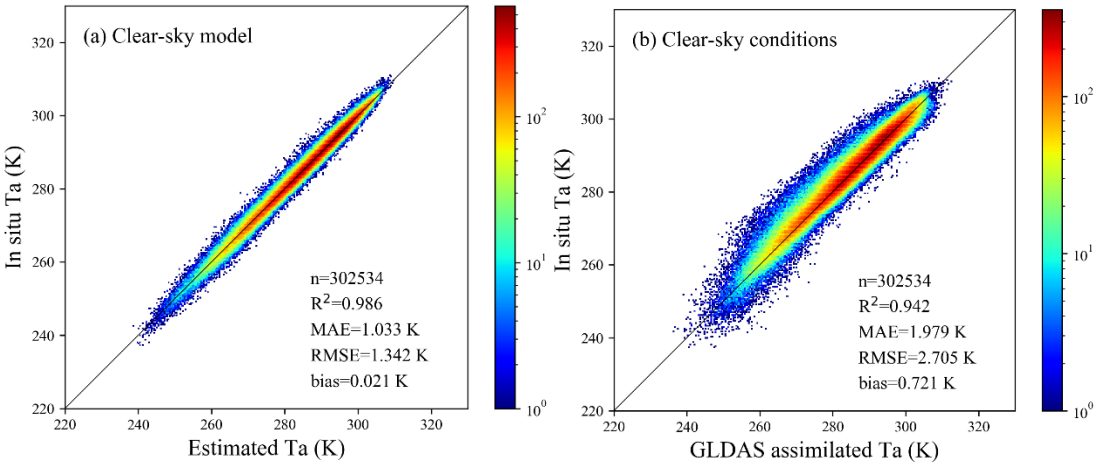


Figure 2. The RMSE histogram of stations in southwest China.

Point 2: In section 3.2, GLDAS assimilated Ta was used in three models as input features. In the feature importance of those three models, assimilated Ta ranked first for two cloudy models and was second to Terra nighttime LST for clear-sky models. The second biggest concern for this study is that it seems like GLDAS assimilated Ta determines the RMSE and R^2. From the fourth paragraph in introduction part, no author used assimilated Ta as the predictor. Instead, shen [1] only used the soil mositure content, albedo and soil evaporation from GLDAS as predictors. If the ground-based Ta ingested by GLDAS was introduced as the preditor, whether it is a circular reasoning that reach better results? I would suggest removing the assimilated Ta as the predictor for three models.

Response 2: Thank the reviewer for making the valuable comments. Since the GLDAS assimilated T_a has well captured the spatial and temporal variation of the actual T_a, it is not surprising to see the great contributions of the GLDAS T_a. However, GLDAS T_a does not completely determine the RMSE and R^2 of our models because many additional inputs have greatly improved the T_a prediction. As shown in Fig. 3, the RMSE values of the GLDAS T_a under three weather conditions are 2.705 K, 2.545 K, and 2.588 K, respectively, while our final models have much better results (RMSE values are 1.342 K, 1.440 K, and 1.396 K, respectively).



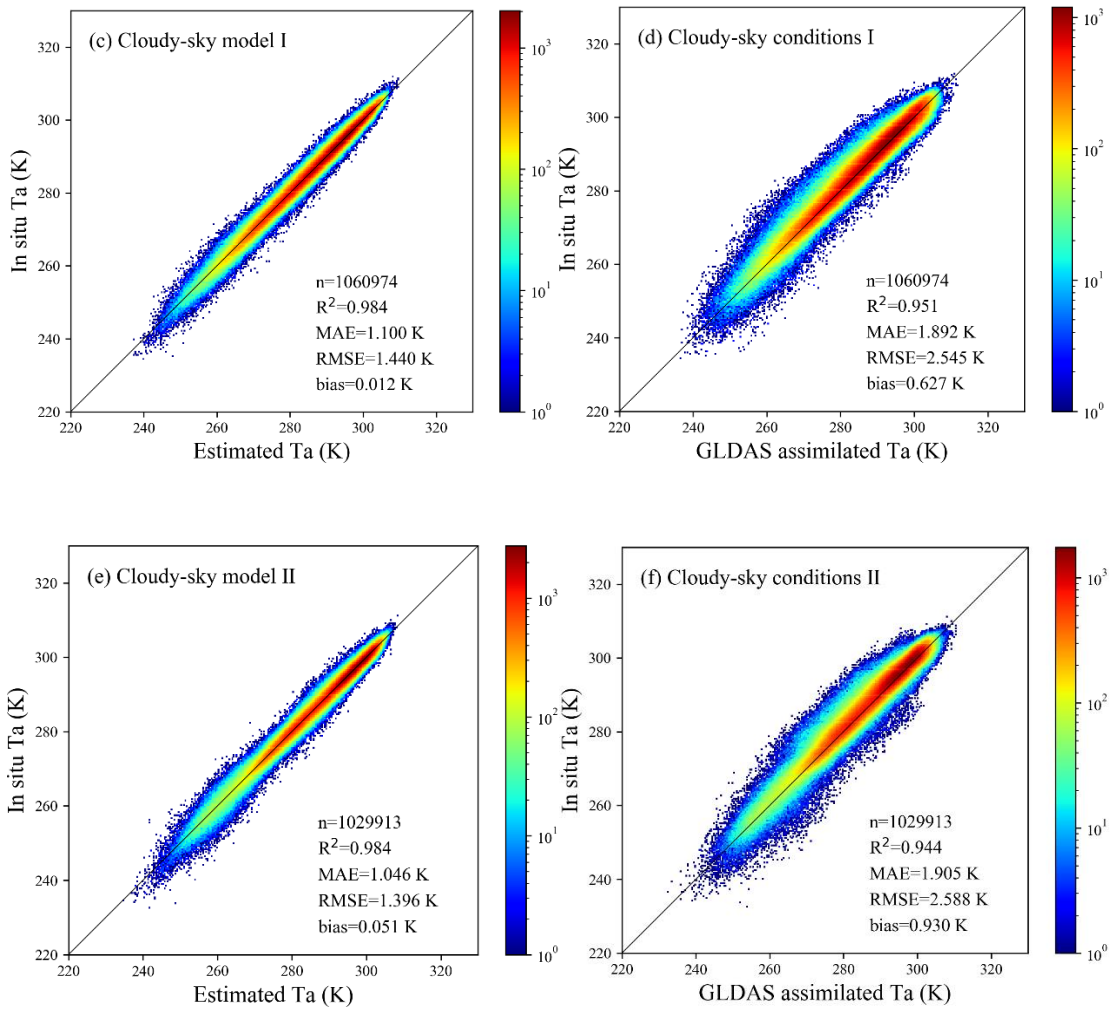


Figure 3. Density scatter plots of the estimated T_a and GLDAS assimilated T_a against the station observed T_a. (a, c, e) are the RF T_a under three weather conditions, (b, d, f) are the GLDAS assimilated T_a under three weather conditions.

Before conducting this study, we did read the paper of Shen et al. (2020) carefully and conducted some experiments. We believe, also based on our initial experiments, that use of GLDAS T_a as a predictor is a much better choice than GLDAS soil moisture (SM), albedo and evaporation because GLDAS assimilated a huge amount of T_a observations into the model to "control" the calculated T_a, while SM, albedo and evaporation are calculated outputs and have much larger uncertainties. The predictors need to be as accurate as possible.

Incorporating GLDAS T_a as our model predictor is not a circular reasoning issue since GLDAS T_a can be considered to be a priori knowledge. Use of a priori knowledge has been the common practice in quantitative remote sensing (Liang, 2004; Liang and Wang, 2019).

In fact, after removing GLDAS T_a as the predictor, the validation statistics of the three models are worsened as shown in Fig. 4, especially for cloudy-sky model II, which does not include MODIS LST at all. RMSE values of the three models were 1.498 K,

1.859 K, and 2.359 K, respectively, which increased by 0.156 K, 0.419 K, and 0.963 K compared with that before removing GLDAS T_a, respectively. It proved that GLDAS T_a was used as a priori knowledge in this study, rather than completely determining the prediction results. Therefore, we still keep GLDAS T_a as the predictor.

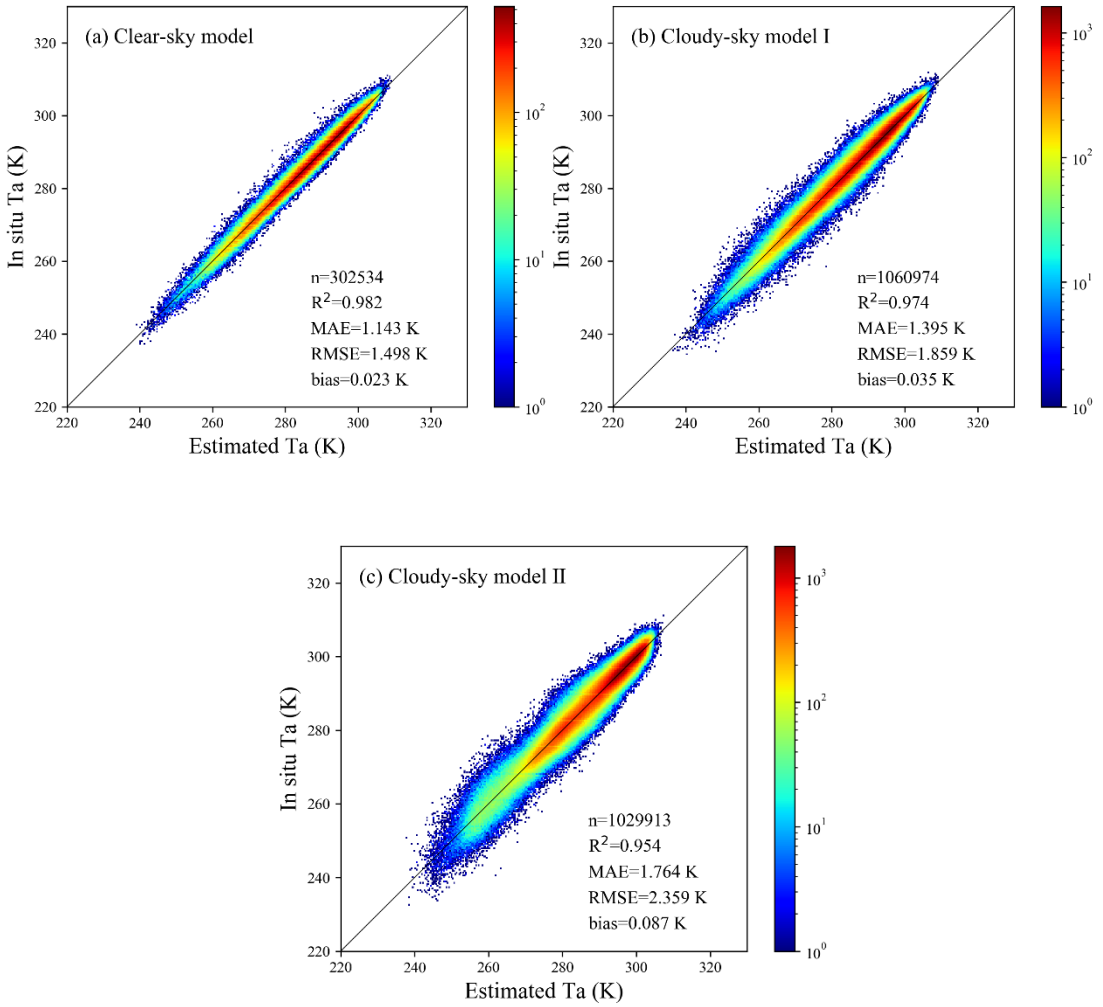


Figure 4. Density scatter plots of the T_a estimated by the models with GLDAS T_a removed against the station observed T_a.

Point 3: The smallest concern is the spatiotemporal model validation strategy in this study which just relys on random cross validation.
However, ignoring spatial and time dependence in model cross-validation can create false confidence in model predictions and hide model overfitting, and this problem that has been well documented in recent works [2, 3]. Please give explanations why this study still used an overoptimistic approach (random cross validation) to assess the prediction error in both space and time.

Response 3: Thank you for your nice comments. To test the models' performance in predicting conditions beyond the temporal and spatial location of the training data, we further used the two validation strategies of Leave-Time-Out (LTO) cross-validation

(CV) and Leave-Location-Out (LLO) CV on the basis of random sample validation. These two strategies have been used in some studies to evaluate the performance of spatiotemporal models in unknown time or unknown space (Liu et al., 2020; Ploton et al., 2020; Xiao et al., 2018).

First, for LTO CV, we divided the data pairs from 2003 to 2016 into 14 groups by calendar year. In each iteration, 13 groups of data were used as training set for model training, and the remaining one group of data was used for validation. The modeling and validation process were repeated 14 times until each year's data was validated. The results are shown in Fig. 5. The RMSE values of validation results for different groups of data range from 1.359 K to 1.665 K. The minor difference between the LTO CV results proves that these models have good extensibility in time.

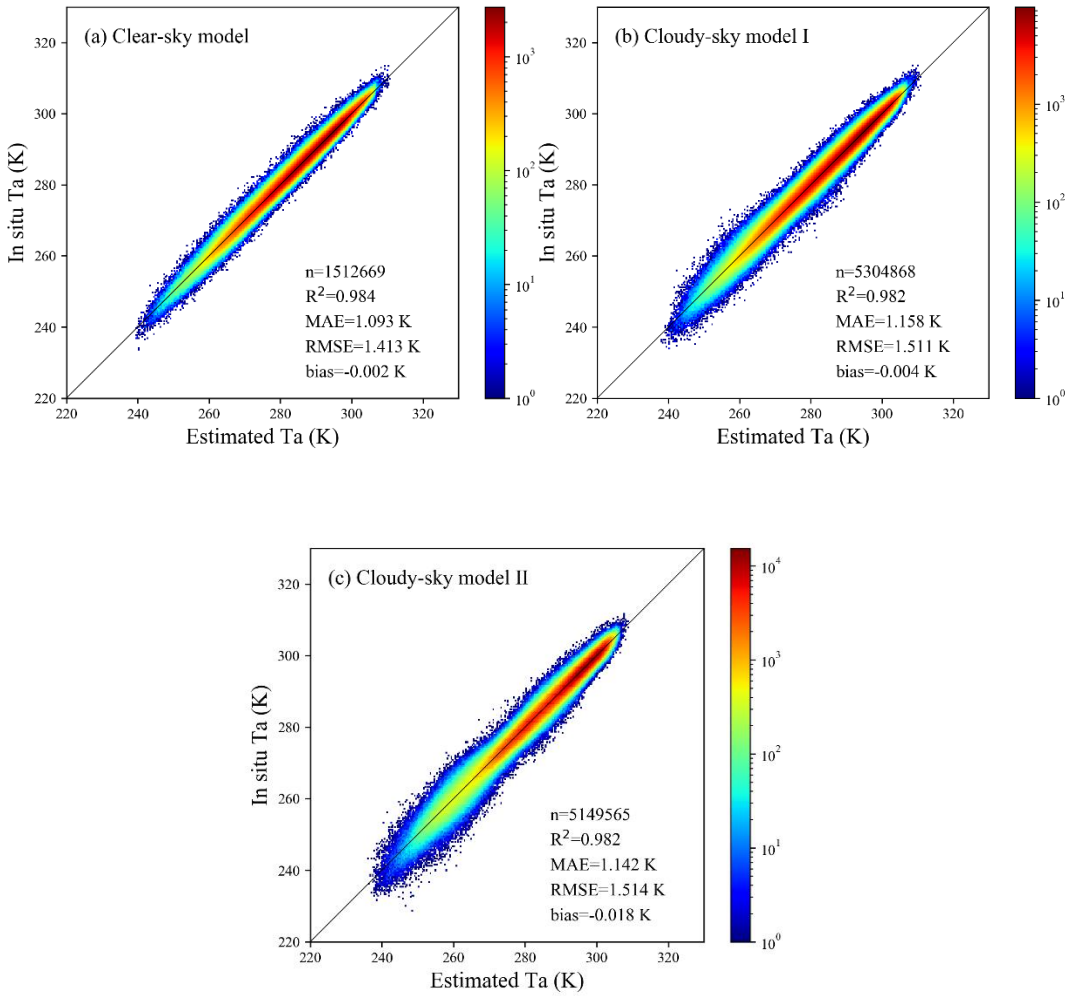


Figure 5. Density scatter plots of LTO CV results for three models.

Then, for LLO CV, we divided 7 clusters in the Chinese region as shown in Fig. 6 by using the similar separation strategy of Xiao et al. (2018). Stations used in this study were divided into different clusters according to their spatial locations, and all data pairs were divided into 7 groups according to the cluster of station. In each iteration, 6 groups of data were used as training set and the remaining one group of data was used for

validation. The modeling and validation process were repeated 7 times until the data of each group was validated. The total validation results of the models under three weather conditions are shown in Fig. 7, with RMSE values ranging from 1.615 K to 1.957 K. As expected, the prediction error of LLO CV increased relative to random sample validation. This is because the relationship between T_a and other features varies with geographical location. The prediction error of the Northwest and Southwest clusters was larger than that of other clusters. RMSE values of these two clusters exceeded 2.5 K under cloudy-sky conditions II while RMSE values of the other clusters were about 1.5 K. This is consistent with the analysis of the spatial distribution of model accuracy in section 4.4 of the manuscript. The meteorological stations in Northwest China and the Qinghai-Tibet are distributed discretely and far away from other stations in China, leading to a large difference between the training set and the test set, and ultimately resulting in the relatively poor performance in the LLO CV strategy in these two regions. Furthermore, the LLO CV results of the cloudy-sky model II are worse than those of the clear-sky model and cloudy-sky model I, indicating that LSTs help to reduce the spatial overfitting of the models.

We have added the content on page 13-14, lines 275-284 and page 19-21, lines 370-397 in the revised manuscript:

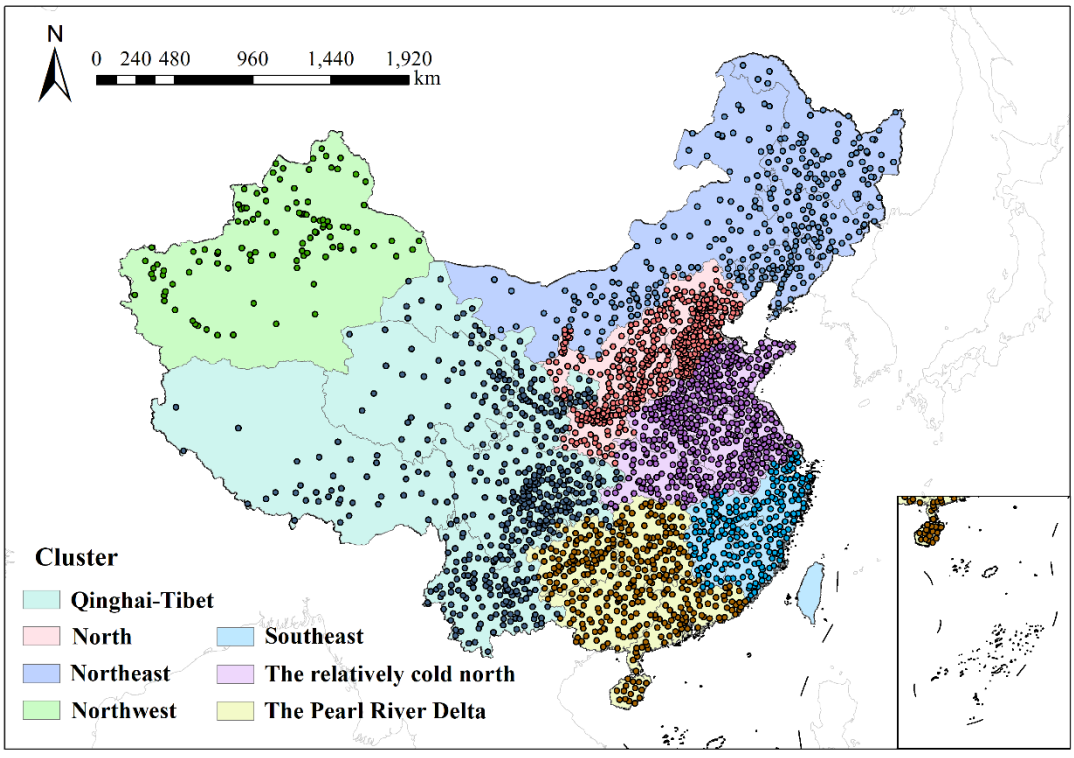


Figure 6. Cluster separation in the research area. According to geographical distribution, mainland China is divided into 7 clusters, which are the North, the Northeast, the Northwest, the Southeast, the relatively cold north, the Qinghai-Tibet Plateau, and the Pearl River Delta, respectively

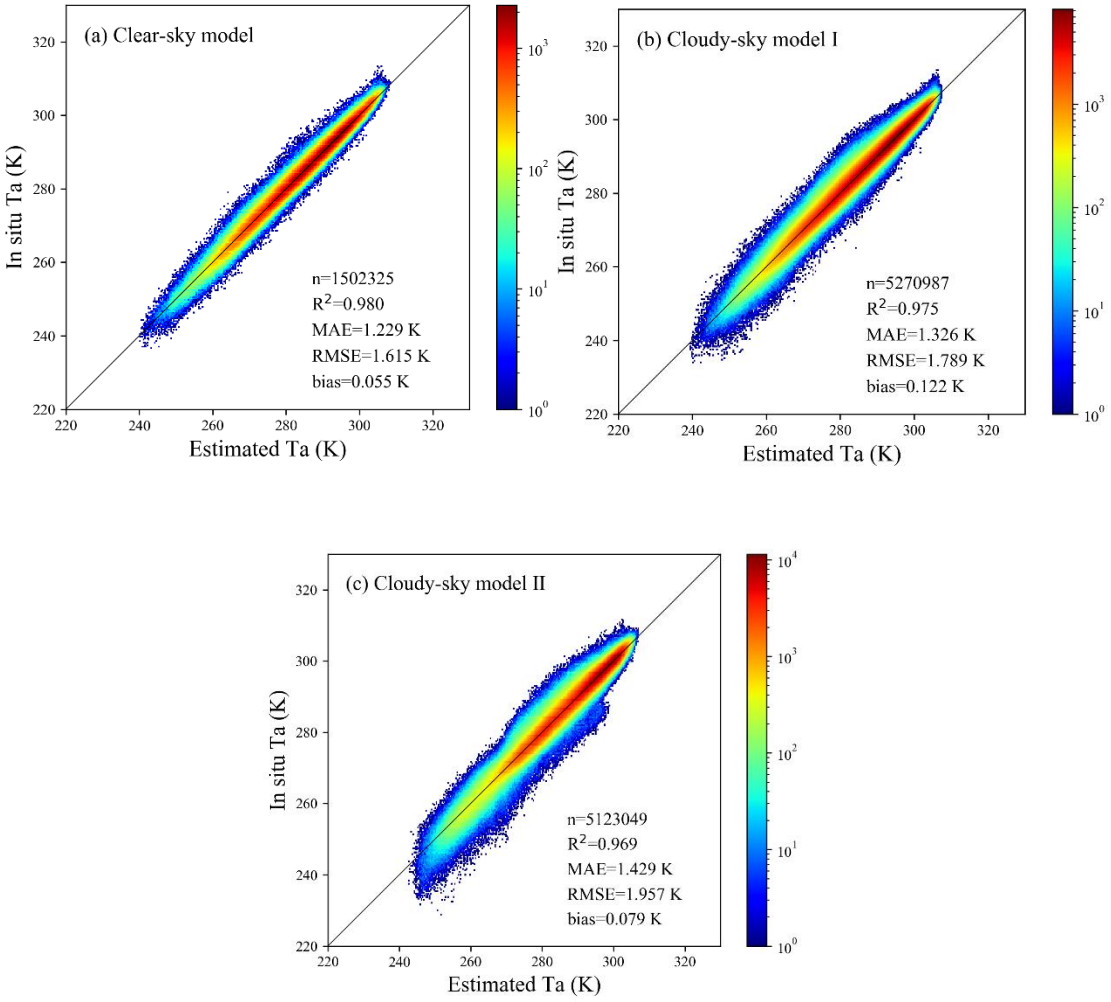


Figure 7. Density scatter plots of LLO CV results for three models.

The T_a predicted by the models was compared with the corresponding station observations. RMSE, MAE, and R^2 were selected as criteria for model evaluation. In order to comprehensively evaluate the performance of the models, we adopted three model validation strategies: random sample validation, LTO CV, and LLO CV. For random sample validation, test set (1/5 of the total data from 2003 to 2016 selected randomly) was used to evaluate the performance of the final T_a estimation models. The results were grouped by elevation range, land cover type, and month to evaluate the model performance under

different situations. For LTO CV and LLO CV, we divided all data pairs into 14 groups according to calendar year and 7 groups according to geographical location. In each iteration, one group of data was used for validation, and the other groups of data were used as the training set for model training. The modeling and validation process were repeated 14 and 7 times until each year's data and each cluster of data was validated. These two CV strategies have been used in some studies to evaluate the performance of spatiotemporal models in unknown time or unknown space (Liu et al., 2020; Ploton et al., 2020; Xiao et al., 2018). ~~To evaluate the performance of the RF models, the prediction results for the test sets were compared with the corresponding station observations. RMSE, MAE and R^2 were selected as criteria for model evaluation. The results were grouped by elevation range, land cover type, and month to evaluate the model performance under different situations.~~

4.2 Cross-validation

In addition to random sample validation, two CV methods were used to further evaluate model performance. For LTO CV, we divided the data pairs from 2003 to 2016 into 14 groups by calendar year. In each iteration, 13 groups of data were used as training set for model training, and the remaining one group of data was used for validation. The modeling and validation process were repeated 14 times until each year's data was validated. The results are shown in Fig. 8. The RMSE values of validation results for different groups of data ranged from 1.359 K to 1.665 K. The minor difference between the LTO CV results proved that these models have good extensibility in time.

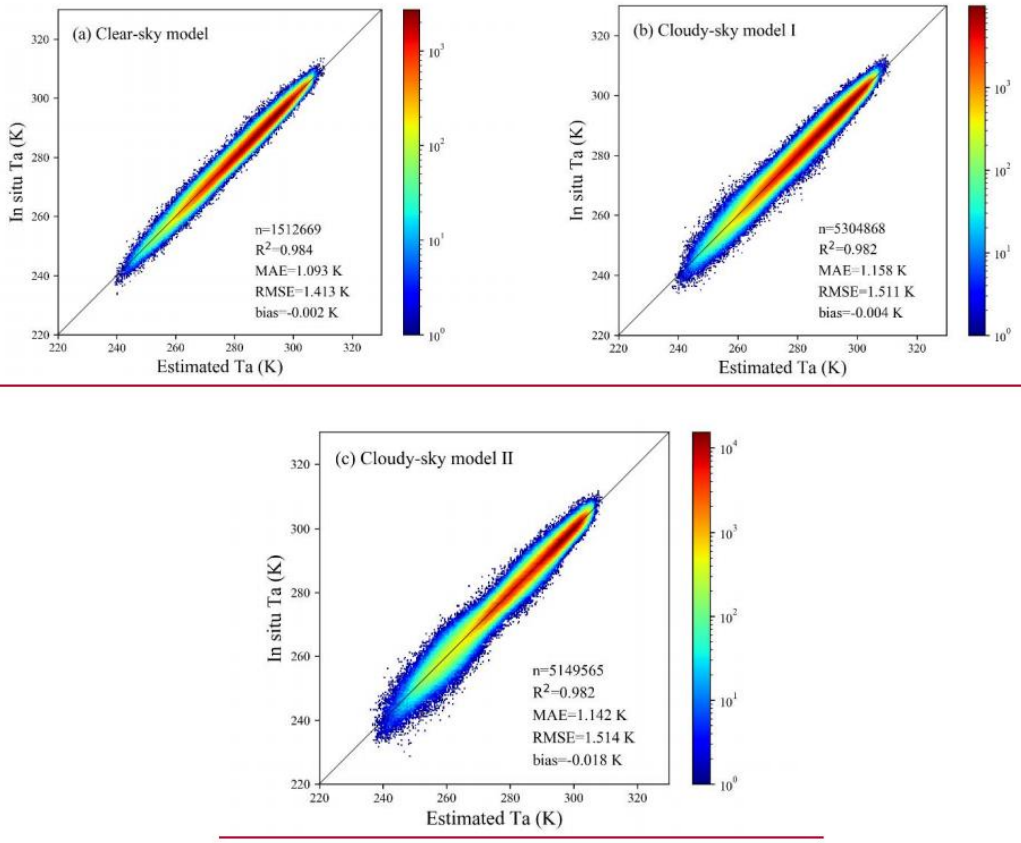


Figure 8. Density scatter plots of LTO CV results for three models.

Then, for LLO CV, we divided 7 clusters in the Chinese region by using the similar separation strategy of Xiao et al. (2018). Stations used in this study were divided into different clusters according to their spatial locations, and all data pairs were

divided into 7 groups according to the cluster of station. In each iteration, 6 groups of data were used as training set and the remaining one group of data was used for validation. The modeling and validation process were repeated 7 times until the data of each group was validated. The total validation results of the models under three weather conditions are shown in Fig. 9, with RMSE values ranging from 1.615 K to 1.957 K. As expected, the error of LLO CV increased relative to random sample validation. This is because the relationship between T_a and other features varies with geographical location. The prediction error of the northwest and southwest clusters was larger than that of other clusters. RMSE values of these two clusters exceeded 2.5 K under cloudy-sky conditions II while RMSE values of the other clusters were about 1.5 K. This is consistent with the analysis of the spatial distribution of model accuracy in section 4.4 of the manuscript. The meteorological stations in northwest and southwest China are distributed discretely and far away from other stations in China, leading to a large difference between the training set and the test set, and ultimately resulting in the relatively poor performance in the LLO CV strategy in these two regions. Furthermore, the LLO CV results of the cloudy-sky model II were worse than those of the clear-sky model and cloudy-sky model I, indicating that LSTs help to reduce the spatial overfitting of the models.

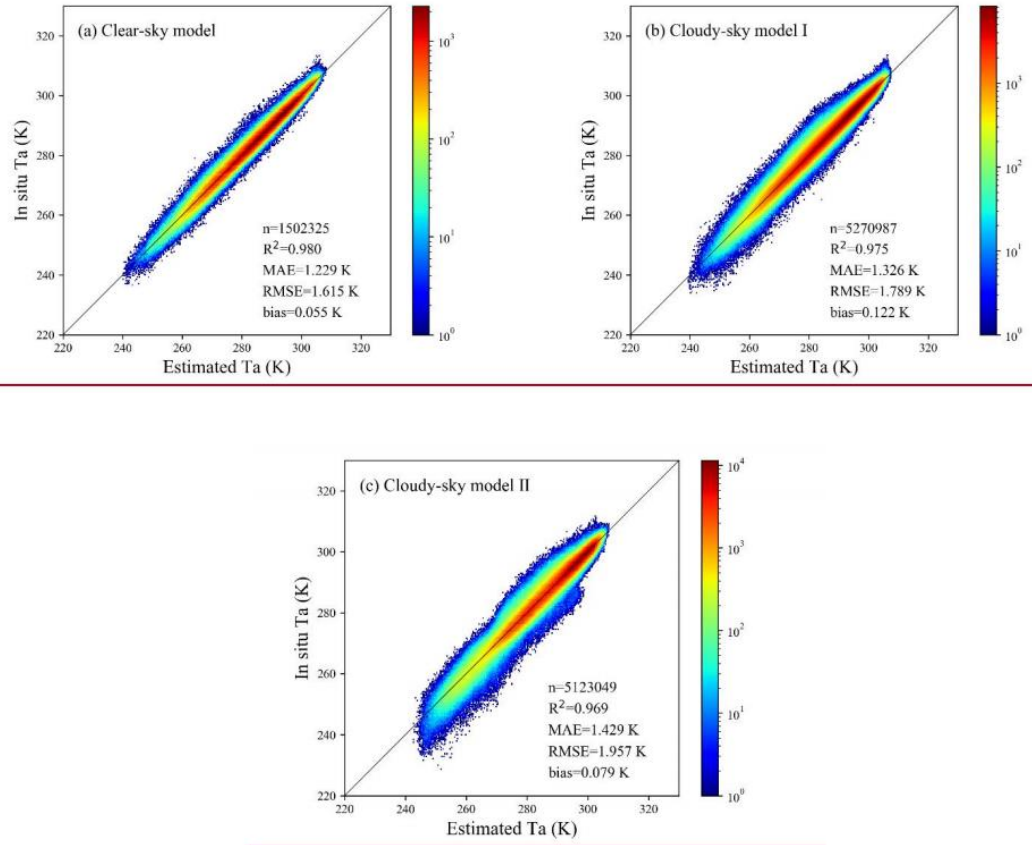


Figure 9. Density scatter plots of LLO CV results for three models.

References:

Liang, S.: Quantitative remote sensing of land surfaces, John Wiley & Sons, Inc., Hoboken, NJ, USA, 2004.

Liang, S., and Wang, J.: Advanced remote sensing: terrestrial information extraction and applications, 2 ed., Academic Press, 2019.

Liu, R., Ma, Z., Liu, Y., Shao, Y., Zhao, W., and Bi, J.: Spatiotemporal distributions of surface ozone levels in China from 2005 to 2017: A machine learning approach,

Environ. Int., 142, 105823, https://doi.org/10.1016/j.envint.2020.105823, 2020.

Ploton, P., Mortier, F., Rejou-Mechain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., Lyapustin, A., Gourlet-Fleury, S., and Pelissier, R.: Spatial validation reveals poor predictive performance of large-scale ecological mapping models, Nat. Commun., 11, 4540, https://doi.org/10.1038/s41467-020-18321-y, 2020.

Shen, H., Jiang, Y., Li, T., Cheng, Q., Zeng, C., and Zhang, L.: Deep learning-based air temperature mapping by fusing remote sensing, station, simulation and socioeconomic data, Remote Sens. Environ., 240, https://doi.org/10.1016/j.rse.2020.111692, 2020.

Xiao, Q., Chang, H. H., Geng, G., and Liu, Y.: An Ensemble Machine-Learning Model to Predict Historical PM2.5 Concentrations in China from Satellite Data, Environ. Sci. Technol., 52, 13260-13269, https://doi.org/10.1021/acs.est.8b02917, 2018.