Reviewer #1

I have checked the response submitted by the author, and most problems have been solved. But I still have some problems to solve before the manuscript is accepted and published (Minor revision)

Comment #1-1: The overlapping region was solved by the authors in the revised manuscript, but why the statistical accuracy in Table 2 did not change, especially in the South America region? According to the author's description, the model of each region is independent. Compared with the original paper, the number of stations in the South America region is significantly reduced, but the statistical results of Tmax and Tmin in South America in Table 1, Figure 4, and Figure 5 are not changed, which makes me very confused.

Response: Thank you very much for your comments. There are some changes related to South America. Accordingly, we have updated Tables 1, S1, S2, Figures 4, 5, S1, S3, and S5 at the regional level in the revised manuscript and supplement. However, the statistical accuracy in Tables 2 – 4 at the global level did not change, because there are only a few small changes in South America and the number of records for validation in South America are very small compared to the global total.

Comment #1-2: I don't agree with the response to my comment 9. I don't think the 10-fold cross validation method in stability has superiority over independent validation. Although the training and test sets of 10-fold cross validation are independent, the correlation between neighboring stations is often ignored for regions with a high density of ground-based stations, so the difference between the results of 10-fold cross validation and random sampling validation is small in regions with high station density (e.g., North America). At the same time, there are some differences in the validation results in South America and Australia, where the station density is relatively small. Given the spatial correlation between stations, the independent validation methods for different years of data as training set and validation respectively (such as the data from 2003-2018 as the training set, and the data in 2020 as the validation set) can more reasonably portray the retrieval accuracy of Tmax and Tmin in the pixel region of the missing ground-based stations.

Response: Thank you very much for the suggestion. We used the 10-fold cross-validation in this study for two main reasons. First, 10-fold cross-validation is more reliable compared to the random sampling validation (Figure R1), especially for regions with a low station density. Both methods were implemented based on independent random sampling and do not have large differences except for the number of evaluations. In this experiment (Figure R1), the result shows that the accuracies vary largely across five evaluations using the random sampling method. The differences in accuracy between the two validation methods and also across the evaluations using the random sampling method were mainly caused by their sensitivities to station densities. The accuracy assessment results are more stable in regions with a higher station density (e.g., North America) because more stations were used in both the 10-fold cross-validation and the suggested random sampling method. Compared to the random sampling method, our 10-fold cross-validation is more reliable, especially in regions with a low station density.

Second, the suggested validation method using data from a different year is not applicable to our method in this study because spatial correlations between Ta and explanatory variables were not assumed constant across days and years. We agree that some models (e.g., random forest and deep learning) in literature have implemented temporal evaluations of model performance due to their underlying assumptions on fixed spatiotemporal correlations. However, in this study, our model is similar as a spatial interpolation technique.

It is not applicable to assume the same spatial correlations between Ta and explanatory variables across time. For example, there is no theoretical basis to find two corresponding days from different years as training and testing days. Therefore, we fitted the SVCM-SP model and estimated the gridded Ta using data in the same day.

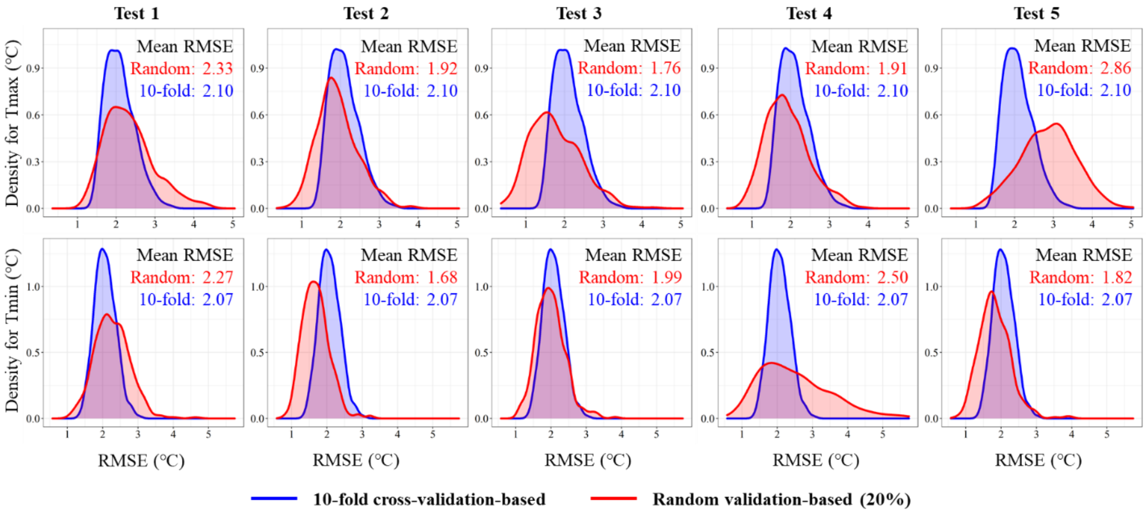

Figure R1: Comparison of RMSEs (top: Tmax; bottom:Tmin) in Africa in 2010 between the five evaluation using the random sampling method (red) and 10-fold cross-validation (blue).

Comment #1-3: The regression line in Figure 3 is incorrect, for example, the regression equation between the estimated and measured values of Tmax over the African region is Y=1X, which means that the estimated Tmax and measured Tmax are exactly the same, which is obviously inconsistent with the value of R (not equal to 1) and the scatter plot, so the form of the regression equation should be modified to "Y=aX+b"

Response: Thank you for your suggestion. As suggested, we have revised Figure 3 below.

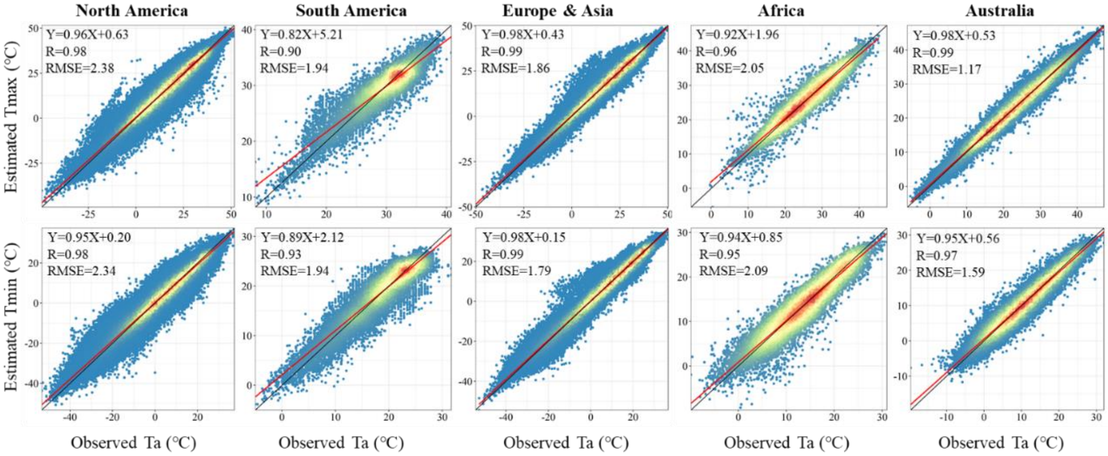

Figure 3: Scatter plots between estimated and observed Ta in five regions in year 2010. Each point represents the estimated and observed Ta (Tmax or Tmin) in a specific day in a weather station. The color of points represents the density, in which red and blue points represent the high and low densities, respectively. The red line is the regression line and the black line is the 1:1 line.