Reviewer #1

Comment #1-1: Satellite remote sensing and ground-based stations play their unique advantages in global meteorological parameter retrieval, and near-surface temperature is of great significance to global and local climate. In this study, satellite and ground-based data are used to retrieve near-surface maximum and minimum temperatures in almost global land regions with 1-km resolution. Overall, this manuscript is clear and well written and presents interesting research, however I found there was a lack of details necessary to fully understand the methods. Some concerns are needed to address, below are my major comments. (Major revision)

Response: Thank you very much for your suggestions. We have provided more details to clarify our method. We first clarified the similarities and differences between current study and the study by Zhang et al. (2022b). We then explained the time difference between LST and Ta, the suitability of 10-fold cross-validation, improved figures, and also added scatter plots for validation of the retrieval results. Below please find our responses to your comments in detail.

Comment #1-2: Line 16: I think "ground station-based Ta" is better than "station-based ground Ta".

Response: Done!

Comment #1-3: The descriptionof the "global dataset" is not rigorous, as the authors have only achieved retrieval of near-surface temperatures for most of the global land region.

Response: Thank you for your suggestion. We really appreciate your point. We keep 'global dataset' in the title to follow the conventions in the community. For example, the famous CRU dataset (https://www.ipcc-data.org/observ/clim/cru_climatologies.html) is one of the best global climate datasets although it is land-only. But we changed the title to 'A global dataset of daily maximum and minimum near-surface air temperature at 1-km resolution over land (2003-2020)' in the revised manuscript as suggested.

Comment #1-4: Lines 82-83: Different regional retrieval models have certain differences. Two different Tmaxs and Tmins will be obtained in the overlapping region. How does the author calculate the final Tmax and Tmin results?

Response: Thank you for your question. There is a significant overlapping region between North America and Latin America in our dataset. We agreed that different Tmaxs and Tmins can be obtained in the overlapping region. In the revised dataset, we removed the largely overlapping region between Latin America and North America to help users to use the dataset.

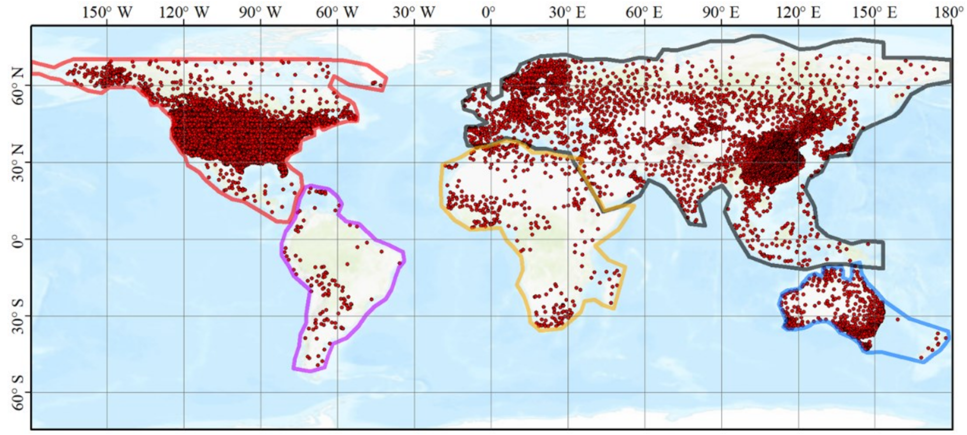


Figure 1: Regions and locations of weather stations in this study. Red points are the locations of weather stations, polygons are the boundary of regions used in the SVCM-SP algorithm. Specifically, polygons of red, purple, orange, blue, and black represent the boundaries of North America, South America, Africa, Australia, and Europe & Asia, respectively.

Comment #1-5: The latitude and longitude grid should be added in Figure 1.
Response: We have improved the figure in the revised manuscript as suggested.

Comment #1-6: Please explain in detail the similarities and differences with the study by Zhang et al. (2022b), and whether there are other differences besides the study area.
Response: Thank you for your question. The study by Zhang et al. (2022b) focuses on the development of the SVCM-SP algorithm for estimating gridded Ta, taking mainland China as an example. It contains many details on the SVCM-SP algorithm and was systematically compared with the geographically weighted regression (GWR) for the novelty of the method. As ESSD is interested in the publication of articles on original research data (sets), this paper focused on the characteristics (e.g., accuracy, spatial and temporal patterns) of the original gridded Ta dataset using the developed and validated SVCM-SP algorithm from Zhang et al. (2022b). Research papers are not within the scope of ESSD. It contains meticulous model calibration and accuracy assessment on the data product in different continents, land cover types, elevation ranges, and climate zones. It also includes several examples on the spatial and temporal distribution of Ta data. Besides, the advantages of the gridded Ta were shown by comparing with existing gridded Ta datasets. We added descriptions in the revised manuscript and removed replicated texts between the two studies.

"Zhang et al. (2022b) successfully estimated and validated gridded Ta using the SVCM-SP algorithm and demonstrated its novelty through the comparison with the geographically weighted regression (GWR) model, while in this study, we developed the global product of gridded Ta, performed extensive model calibration and accuracy assessment at the global scale, and provided details on accuracy, spatial and temporal patterns of the global gridded Ta." (lines 76-79)

Comment #1-7: Lines 107-114: The ground-based stations usually have a high temporal resolution, and the seamless global surface temperature data used by the author comes from the MODIS sensor, which has a limited number of transits. When the training sample library is constructed, the author only describes the spatial matching process and ignores the temporal matching, please give a detailed description.

Response: Thank you for your suggestion. Mid-daytime and mid-nighttime LSTs were used to develop their relationship with air temperature to interpolate station Tmax and Tmin, respectively. We agree that there may be time difference between Ta and LST used in the algorithm. The key information about LST in our algorithm is its spatial variations. Within the small difference in time between LST and Ta, there will not be significant change in the spatial variations of LST. Therefore, the impact of time difference between LST and Tmax/Tmin on the accuracy of the estimated Ta is minor. We proved it by adding shifts to LST for estimating Ta in North America in an example (Fig. S8). We added discussion in the revised manuscript.

"Specifically, mid-daytime and mid-nighttime LSTs were used to develop their relationship with air temperature to interpolate station Tmax and Tmin, respectively. The actual time of Tmax and Tmin may be slightly different from mid-daytime and mid-nighttime of LST. Within the small difference in time between

LST and Tmax/Tmin, there will not be significant change in the spatial variations of LST. Therefore, the impact of time difference between LST and Tmax/Tmin on the accuracy of the estimated Ta is minor as shown by shifting LST for time difference (Fig. S8)." (lines 118-123)

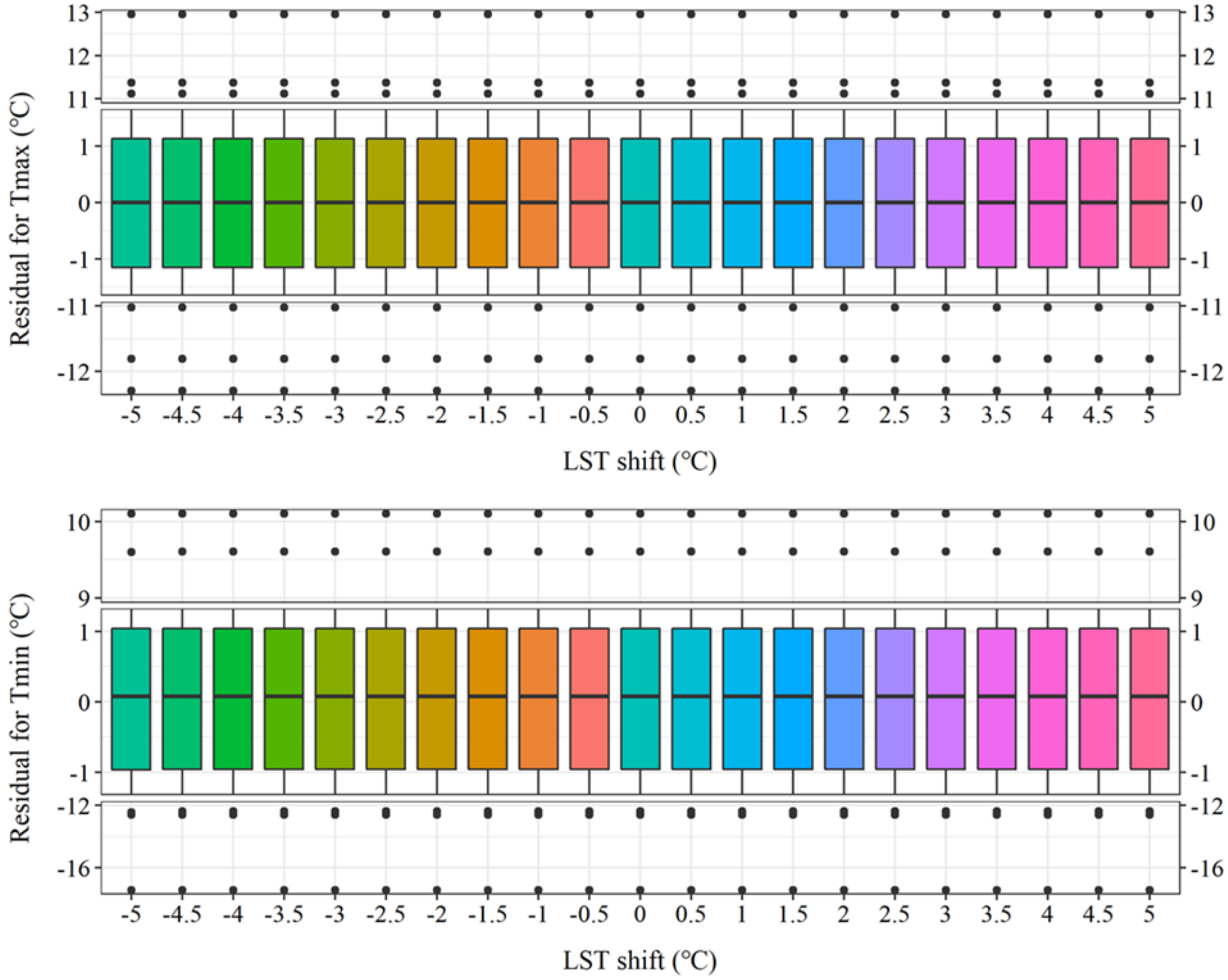


Figure S8: Residuals of estimated Ta based on the 10-fold cross-validation by shifting LSTs from -5 to 5 ℃ at a step of 0.5 ℃ in North America in the day 200 of 2010. Each box is based on the residual between observed and estimated Ta in the stations. Each point represents the residual in a specific station.

Comment #1-8: Lines 124-125: Does the authors mean that the ground-based measurement result corresponding to mid-daytime is the Tmax of the site, and the ground-based measurement result corresponding to mid-nighttime is the Tmin of the site? If yes, I don't think it's reasonable, especially for the Tmin.

Response: Thank you for your question and comment. As we explained in the response to your comment #1-7, we did not equate mid-daytime LST to Tmax or mid-nighttime LST to Tmax. Instead, our models are empirical/correlative in nature, and mid-daytime (or mid-nighttime) LST is correlated strongly with the true Tmax (or Tmin), which is the theoretical basis for our method as well as for all the existing methods in the literature. We agree that there may be time difference between Ta and LST used in the algorithm. But the impact of such differences in time on the interpolated Tmax/Tmin is minor. Please see details in our previous response.

Comment #1-9: Due to the spatial correlation of the near-surface air temperature at the different stations and the temporal correlation between the training data at same station, the 10-fold cross-validation verification cannot truly reflect the accuracy of the model. The author needs to give independent verification results, such as the first 15 days of each month as training Set, the data of the last 5 days is used as the validation set or test set, or the data from 2003-2018 is used as the training set, and the data in 2020 is used as the validation or test set, or 80% of the site data is used as the training set, and the data from the remaining 20% sites is used as the validation or test set.

Response: Thank you for your suggestion. Our cross-validation is a very rigorous one and the uncertainty estimates are based on independent data (not used for the model training). More importantly, the RMSE represents a conservative estimate of the true uncertainties of our data because when producing the final results, we use all available data, more than those in the 10-fold cross-validation. In order to evaluate the model performance for estimating gridded Ta, we used the widely used 10-fold cross-validation for each day in each region. That is, we equally and randomly divided the valid records into 10 groups. Nine groups were used as training set and the rest one group was used as testing set. This approach was implemented for 10 times until all the groups had been used as testing set. Each test of the 10-fold cross-validation can obtain a RMSE and the average RMSE of the 10 tests was used as the final RMSE. Therefore, the accuracy assessment of the 10-fold cross-validation was implemented based on independent validation data and can provide a reliable evaluation of the accuracy. We compared daily RMSEs based on 10-fold cross-validation and validation with 30% randomly selected testing data (Fig. R1). We found that the two validation methods show similar results in accuracy. The distributions of RMSEs using the 10-fold cross-validation method is more concentrated with higher maximum densities than those of the 30%-random-validation-based method, especially in Africa and South America due to the low number of validation records, indicating the superiority of the 10-fold cross-validation method in stability. We have added relevant descriptions in the revised manuscript.

"The model performance for estimating gridded Ta was assessed based on root mean square error (RMSE) and mean absolute error (MAE) using the 10-fold cross-validation in these regions in each day. Taking the RMSE as an example, a RMSE was generated in each test of the 10-fold cross-validation and all RMSEs from the 10 tests were averaged as the final RMSE in a specific day in a specific region. This accuracy assessment using the 10-fold cross-validation was implemented based on independent validation data and can provide a reliable evaluation of the accuracy." (lines 134-139)

"Specifically, this accuracy assessment represents conservative estimates of the uncertainties of our data because when producing the final results, we used all the available data, more than those in the 10-fold cross-validation." (lines 141-143)

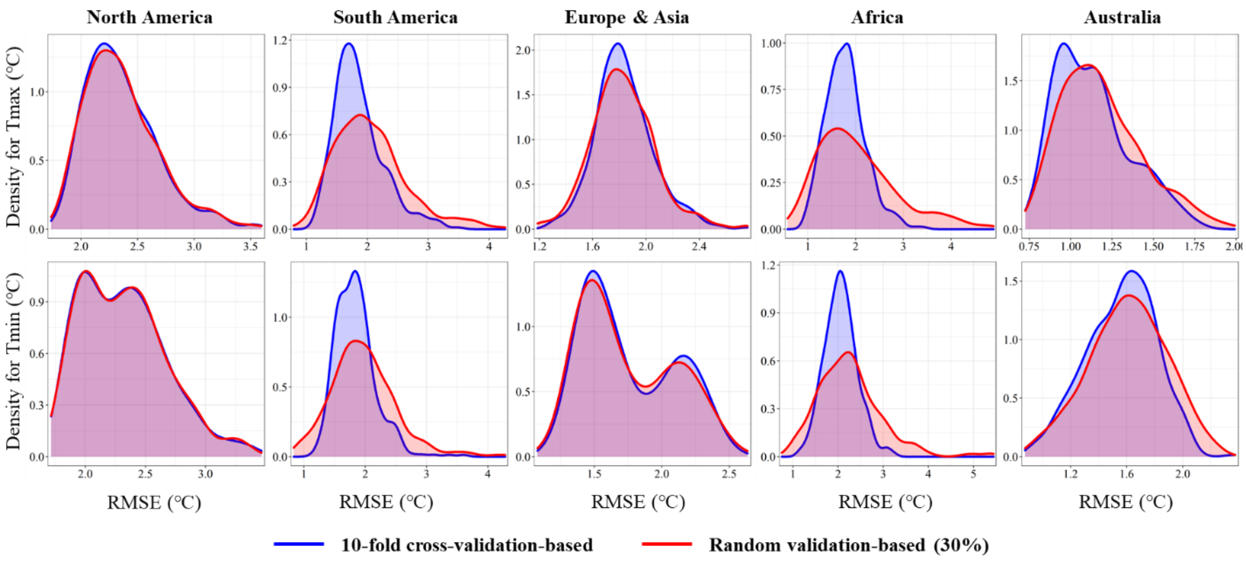


Figure R1: Density of daily RMSEs from 30%-random-validation (red) and mean RMSEs from 10-fold cross-validation (blue) in five regions in year 2010.

Comment #1-10: What does the Y-axis of Figure S1 represent? Is it the number of valid observation samples or the number of stations that only include 1 valid observation sample? Also, I did not find a related description of Figure S1 in the manuscript.

Response: Thank you for your questions. The Y-axis of Figure S1 represents the number of stations with valid records in each day for Tmax (or Tmin). Using the daily number of valid records, we drew a boxplot in each year. We have clarified the descriptions (Figures S1 and S2) in the supplement.

"Figure S1: Number of valid records (Y-axis) in five regions. Each box is based on the daily number of valid records in a specific year. Each point represents the number of valid records in a specific day."

"Figure S3: Number of valid records (Y-axis) after filling missing values in Africa and Latin America. Each box is based on the daily number of valid records in a specific year. Each point represents the number of valid records in a specific day."

"Second, there are missing values, especially in stations in Africa and Latin America (Fig. S1). We filled these data gaps using a 5-day local moving window (Fig. S2). Accordingly, the number of records largely increased (Figs. S3 and S4) with reasonable error ranges (Fig. S5)." (lines 114-117)

Comment #1-11: Comparing Figure 1 and Figure 3 confuses me. In my view, the author has constructed Ta estimation models in 5 study regions, but does not include Greenland (Figure 1). I have three questions. First of all, why can the Ta retrieval in the Greenland region be achieved? The parameters of the retrieval model are the same as in which regions Ta retrieval model? In addition, the standardized regression coefficient for the Greenland region in Figure S7 is also absent. Second, can the constructed retrieval model be used for the retrieval of Ta in the oceanic region? Third, what do the non-color-filled regions on land in Figure 3 (such as the Amazon region and south-central Africa) mean?

Response: Thank you for your questions. (1) Our dataset covers a small portion of Greenland which is constrained by the extent of the global seamless 1 km daily LST dataset. (2) The dataset focuses on the land areas as indicated by the revised title. (3) The non-color-filled regions on land in Figure 3 (Figure 4 in the revised manuscript) are areas without reliable evaluations due to the lack of weather stations. We clarified these in the revised manuscript. As suggested, we improved relevant figures (Figs. 4 and S6) in the revised manuscript and supplement by only showing the regions covered by our dataset.

"Specifically, our dataset covers a small portion of Greenland which is constrained by the extent of the global seamless 1-km daily LST dataset." (lines 102-103)

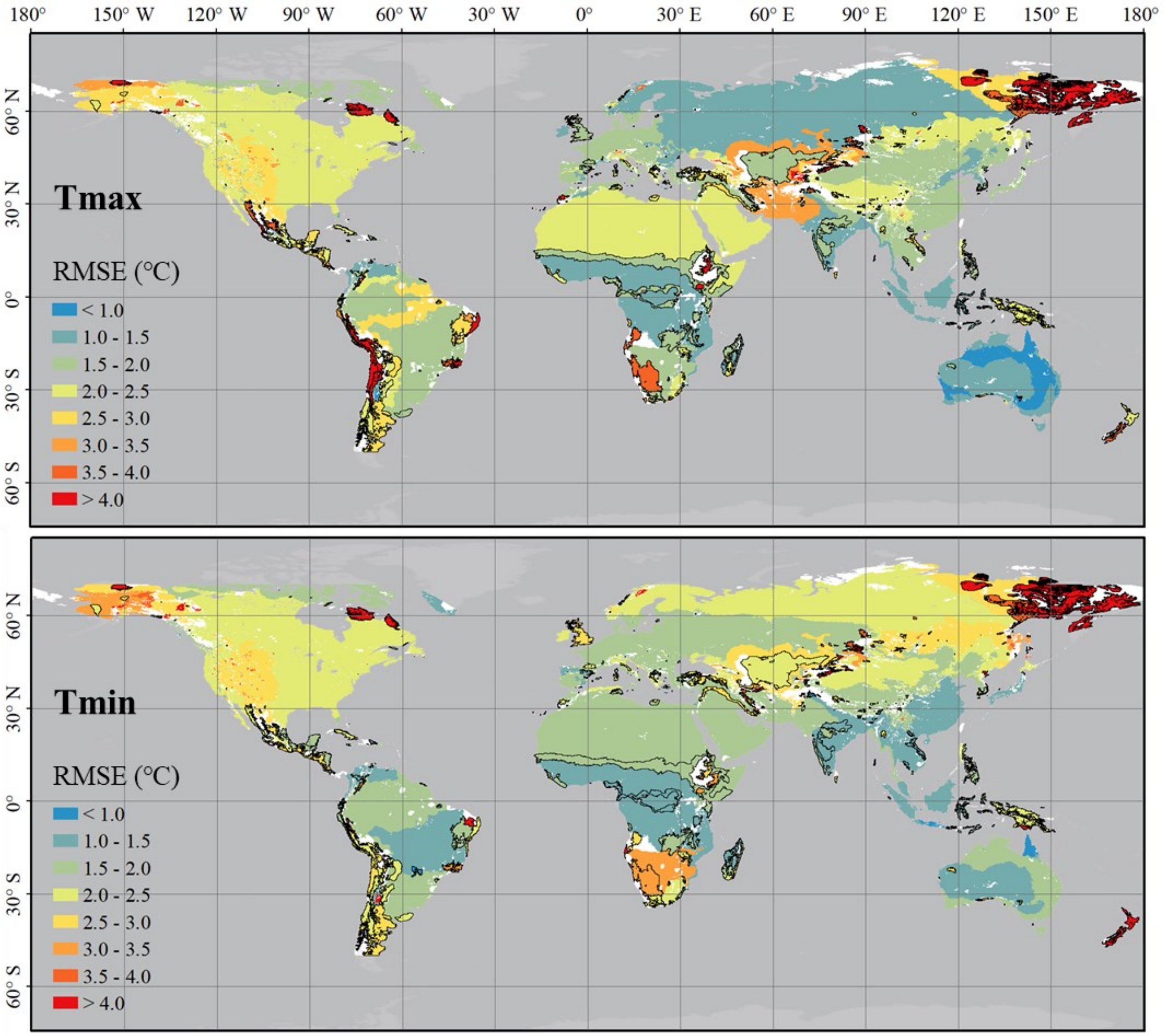


Figure 4: Accuracy of estimated Ta in climate zones in 2003-2020. Climate zones with black boundaries are areas with low densities of weather stations (i.e., distances between training and validation sites are larger than 50 km). The white regions on land are areas without reliable evaluations due to the lack of weather stations. (lines 191-193)

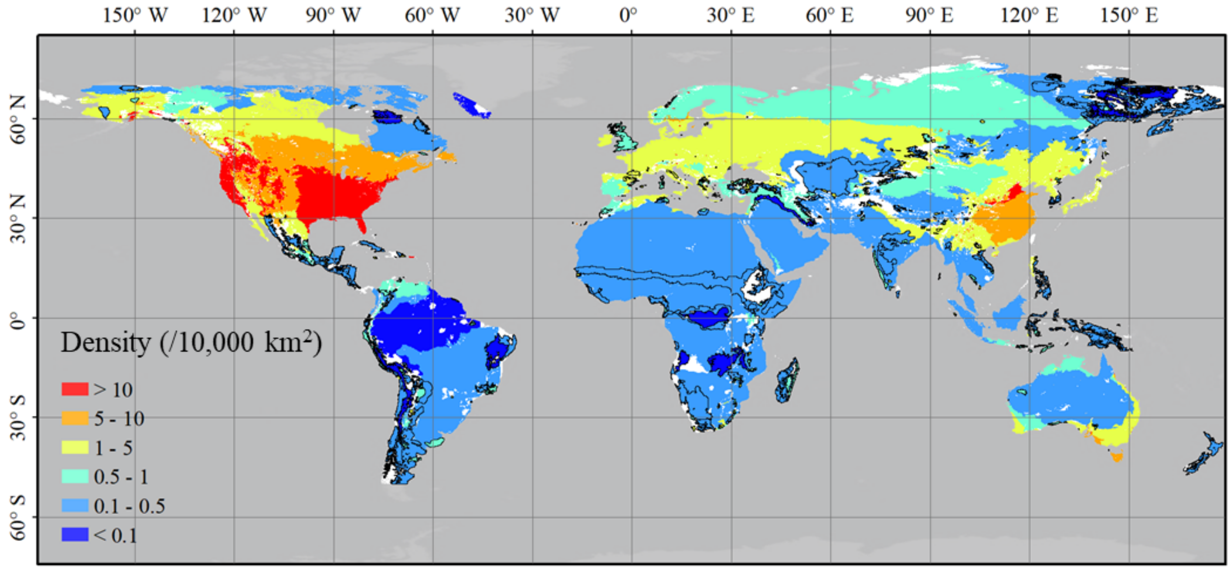


Figure S6: Station density in climate zones in 2003-2020. Climate zones with black boundaries are areas with low densities of weather stations (i.e., distances between training and validation sites are larger than 50 km). The white regions on land are areas without reliable evaluations due to the lack of weather stations.

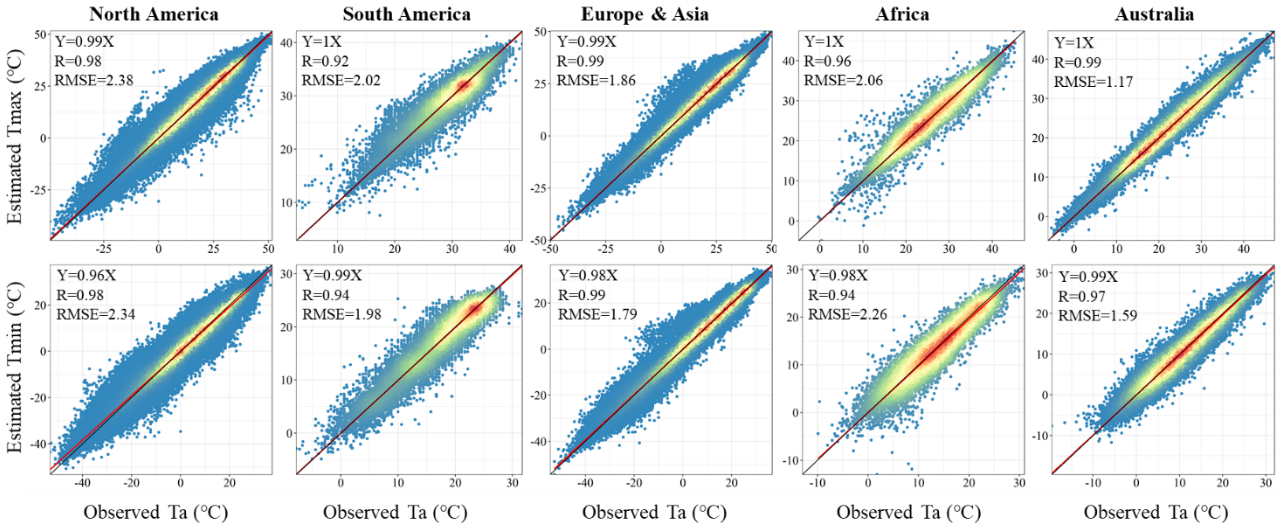
Comment #1-12: It is suggested to modify the color bar range of Tmax and Tmin in Figure 5 to be the same

Response: We improved the figure as suggested.

Comment #1-13: I think that the validation of the results in this manuscript needs to be expanded further and needs to add scatter plots for the validation of the retrieval results, such as density scatter plots instead of just calculating RMSE and MAE.

Response: As suggested, we have added scatter plots in 2010 for validation of the retrieval results.

"The results of the 10-fold cross-validation indicate the accuracy of estimated Ta varies across regions within a reasonable range (Fig. 3 and Table 1). The estimated and observed Ta in different regions scattered along the 1:1 line with the RMSE ranging from 1.17 to 2.38℃ and 1.59 to 2.34℃, respectively, for Tmax and Tmin in 2010 (Fig. 3)." (lines 149-151)



"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." (lines 163-165)